Biostatystyka

– wykorzystanie metod statystycznych w pracy badawczej w naukach biomedycznych

Mojej Rodzinie poświęcam

Biostatystyka

– wykorzystanie metod statystycznych w pracy badawczej w naukach biomedycznych

Cezary Watała

Zakład Zaburzeń Krzepnięcia Krwi
Uniwersytetu Medycznego w Łodzi

α-medica press

ISBN 978-83-7522-072-8

Wydanie II

Korespondencję i uwagi prosimy kierować pod adresem:
prof. dr hab. Cezary Watała
Zakład Zaburzeń Krzepnięcia Krwi, Uniwersytet Medyczny w Łodzi,
ul. Żeromskiego 113, 90-549 Łódź
tel. (42) 63 93 471, fax (42) 67 87 567
e-mail: cezary.watala@umed.lodz.pl
http://www.interhemostaza.pl

Adres wydawnictwa:
poczta@alfamedica.pl
www. alfamedica.pl

Dziękuję wszystkim, których uwagi i sugestie wykorzystałem podczas pisania tej książki. W szczególności, Panu Dziekanowi Wydziału Nauk o Zdrowiu Uniwersytetu Medycznego, prof. dr. hab. med. Tomaszowi Kostce, dziękuję za merytoryczne uwagi, wskazówki oraz dyskusje.
Mojej Żonie Magdzie dziękuję za inspiracje do poznawania nowych obszarów biostatystyki, czerpane z częstych dyskusji na temat stosowania i nauczania metod statystycznych stosowanych w naukach przyrodniczych i biomedycznych.
Czytelnikom wydania I dziękuję za wskazanie mi błędów, pomyłek i niejasności – tylko dzięki ich życzliwym uwagom niektóre z tych braków udało się usunąć.

Spis treści (wersja drukowana)

Spis treści (plik pdf)

Część II Uzupełnienia, przykłady i zadania

Przedmowa do II wydania

Statystyka nie należy do ulubionych przedmiotów, zarówno w czasie studiów, jak i w szkoleniu podyplomowym lekarzy i innych absolwentów studiów medycznych, biologicznych czy humanistycznych. Dotyczy to także tych osób, którzy uczestniczą w ocenie i analizie wyników własnych badań klinicznych, badań skuteczności leków, badań oceniających stan zdrowia populacji czy efekty działań interwencyjnych związanych z profilaktyką rozmaitych chorób czy promocją zdrowia. Wielu uważa, że statystyka jest nudna, zbyt skomplikowana i niezrozumiała, nie tylko dla „zwyczajnych" lekarzy, którzy sporadycznie prezentują swoje dokonania na sympozjach, zjazdach i kongresach, ale także dla pracowników naukowych.

W efekcie powstają liczne publikacje, sprowadzające się do opisu przypadków z własnego oddziału czy kliniki, w których prezentacja danych zawiera proste tabele przedstawiające liczby bezwzględne lub w najlepszym przypadku wartości średnie i odsetki, testowane (o ile w ogóle) testem t-Studenta lub testem chi-kwadrat. Do rzadkości należy pogłębiona dyskusja nad sposobem doboru próby, jej wielkości i schematu badania, czy stosowanie metod wielowymiarowych. Zdaniem świetnego epidemiologa, dr med. Katarzyny Szamotulskiej z Instytutu Matki i Dziecka w Warszawie, konsultacje autora pracy ze statystykiem przebiegają najczęściej według następującego schematu: badacz, niemający żadnej wiedzy w dziedzinie biostatystyki, przynosi ze sobą pendrive'a czy dyskietkę z danymi lub kilkanaście arkuszy papieru, rzuca je na stół i prosi statystyka, niemającego żadnej wiedzy ani doświadczenia w zakresie dziedziny reprezentowanej przez badacza, aby „coś z tym zrobił". Efektem takiej współpracy może być tylko najprostsza, pozbawiona finezji i polotu analiza. Zdarza się także, że badacz decyduje się sam podjąć wyzwanie: kupuje oprogramowanie statystyczne, udaje się na krótki kurs statystyczny i wobec łatwości posługiwania się tym oprogramowaniem, stosuje nieadekwatne narzędzia statystyczne.

W jednym i drugim przypadku efekt końcowy jest podobny: powstają prace niedoskonałe pod względem metodologicznym, które spotykają się z uzasadnioną, krytyczną oceną recenzentów, a często są w ogóle odrzucane, jako niespełniające oczekiwań redakcji krajowych i zagranicznych periodyków. W dobie „medycyny opartej na dowodach" i dużej, stale rosnącej konkurencji w dziedzinie badań naukowych, prace doktorskie, habilitacyjne czy projekty badań zgłaszane do rozmaitych krajowych i zagranicznych instytucji finansujących muszą odpowiadać wysokim standardom metodologicznym. Nie ulega bowiem wątpliwości, że o ocenie publikacji naukowej decyduje nie tylko oryginalność, nowatorstwo i wartość merytoryczna, ale także zastosowanie właściwego schematu badania, poprawność analizy statystycznej i umiejętność interpretacji wyników.

Podobnie, lekarz lub inny pracownik naukowy, zainteresowany krytyczną oceną danych z piśmiennictwa pod względem ich naukowej wiarygodności i przydatności w podejmowaniu konkretnych decyzji w praktyce klinicznej, powinien dysponować niezbędną wiedzą z dziedziny nowoczesnej biostatystyki.

Niestety, na krajowym rynku wydawniczym odczuwa się nadal niedostatek nowoczesnych podręczników – zarówno z dziedziny biostatystyki, jak i epidemiologii. Dlatego uważam, że drugie wydanie „*Biostatystyki*" w opracowaniu prof. Cezarego Watały, zawierające obszerny zbiór metod statystycznych przydatnych w pracy badawczej w naukach

biomedycznych, znakomicie wypełnia tę lukę. Autor, co zapewne zaskoczy niektórych Czytelników, nie jest zawodowym statystykiem czy epidemiologiem, lecz wybitnym specjalistą w dziedzinie biologii medycznej, biologii molekularnej i patofizjologii, a zwłaszcza w zakresie badań dotyczących czynności płytek krwi i układu hemostazy.

Prof. dr hab. Cezary Watała, kierownik Zakładu Zaburzeń Krzepnięcia Krwi Uniwersytetu Medycznego w Łodzi, w przeszłości współpracownik McMaster University w Hamilton (Kanada) i Ohio State University w Columbus (USA), jest autorem ponad 230 publikacji naukowych, w tym ponad 120 prac oryginalnych i przeglądowych opublikowanych w recenzowanych czasopismach zagranicznych o wysokim prestiżu, m.in. *Thrombosis and Haemostasis*, *Thrombosis Research*, *Journal of Molecular Medicine*, *Journal of Laboratory and Clinical Medicine*, *European Journal of Biochemistry*, *European Journal of Haematology*, *Biochimie*, *Current Pharmaceutical Design*, *Biochemical Pharmacology, Diabetes* i wielu innych. Prof. Watała jest kierownikiem, głównym wykonawcą lub współwykonawcą ponad 20 projektów badawczych finansowanych z funduszy unijnych, natowskich lub dotowanych przez Komitet Badań Naukowych (MNiSW). Był wielokrotnie recenzentem prac habilitacyjnych, doktorskich oraz publikacji nadesłanych do krajowych i zagranicznych periodyków. Jest zatem w dziedzinie biostatystyki doświadczonym i pełnym twórczej inwencji praktykiem, czerpiącym wiedzę z najlepszych źródeł. Sądzę, że z tych powodów, jak również ze względu na wysokie kompetencje Autora w dziedzinie nowoczesnej metodologii badań naukowych i dydaktyki medycznej, niniejszy przewodnik biostatystyki, wzbogacony umiejętnie wybranymi przykładami i zadaniami, powinien znaleźć się w bibliotece każdego pracownika naukowego prowadzącego badania w dziedzinie medycyny, biologii lub innych dziedzin związanych z analizą stanu zdrowia, podejmowaniem decyzji klinicznych oraz krytyczną analizą danych piśmiennictwa.

Pierwsze wydanie „*Biostatystyki*" spotkało się z dużym zainteresowaniem i uznaniem, zaś nakład został bardzo szybko wyczerpany. Dobrze zatem, że do dyspozycji wszystkich zainteresowanych trafia drugie – uzupełnione i poprawione – wydanie.

Odbiorcą tego podręcznika może być zarówno młody pracownik naukowy przygotowujący pracę doktorską, doświadczony i biegły klinicysta (niezależnie od stopnia naukowego), jak również student uczelni medycznej stawiający pierwsze kroki w dziedzinie samodzielnych badań naukowych.

Jestem przekonany, że korzyści z tego opracowania mogą odnieść nie tylko ci, którzy myślą o projektach badawczych finansowanych ze środków grantowych Ministerstwa Nauki, Unii Europejskiej lub Fundacji Wellcome, publikacjach w *Nature*, *Science*, *New England Journal of Medicine* lub dobrych, recenzowanych krajowych periodykach naukowych, ale także ci, którzy chcą poznać nowoczesne metody analizy statystycznej i lepiej zrozumieć skomplikowany świat metaanaliz, asocjacji, związków przyczynowo-skutkowych, testów nieparametrycznych czy metod stosowanych w badaniach populacyjnych.

prof. dr hab. med. Wojciech Drygas
Kierownik Katedry Medycyny Społecznej i Zapobiegawczej
Wydziału Nauk o Zdrowiu Uniwersytetu Medycznego w Łodzi
Dyrektor Programu CINDI WHO w Polsce

Łódź, grudzień 2011 r.

Zamiast wstępu...

Niewłaściwie dobrane lub zastosowane metody analizy statystycznej wyników badania naukowego mogą prowadzić do budowania nieprawdziwych wniosków płynących z takiej analizy. Zły dobór metod statystycznych może dotyczyć postępowania analitycznego *a priori* lub *a posteriori*. W pierwszym przypadku, brak właściwego zaplanowania eksperymentu naukowego może doprowadzić do trudności lub nawet niemożności poprawnego opracowania wyników badania. W drugim, kiedy uzyskane wyniki analizujemy przy zastosowaniu niewłaściwie dobranych metod statystycznych, wnioski naszych badań mogą być wypaczeniem lub przekłamaniem rzeczywistości (np. stwierdzając zależności nieistniejące w rzeczywistości).

Niniejsze opracowanie skierowane jest do każdego, kto zajmuje się pracą naukową i w związku z tym ma do czynienia ze statystyczną analizą wyników badań. Zostało ono pomyślane jako wstępne ukierunkowanie i zainteresowanie badacza określonymi procedurami analizy statystycznej, które są powszechnie stosowane przy opracowaniach danych w naukach przyrodniczych, biomedycznych czy farmaceutycznych.

Książka składa się zasadniczo z dwóch części, obejmujących wprowadzenie teoretyczne oraz przykłady i zadania. Bardziej zaawansowani Czytelnicy mogą pominąć teoretyczne wprowadzenie (lub też traktować tą część jedynie pomocniczo) i skupić uwagę wyłącznie na przykładach i zadaniach analizujących wyniki wybranych prac doświadczalnych.

Książka ta nie pretenduje do rangi podręcznika statystyki – jest jedynie zbiorem użytecznych wskazówek, uwag i sugestii na temat tego, jakimi przesłankami powinniśmy się kierować przy wyborze właściwych metod analizy statystycznej, czego unikać, jak przeprowadzić analizę danych w taki sposób, aby ilość uzyskanej informacji była jak największa, oraz jakie niepoprawne wybory mogą nas prowadzić do przekłamania naszych wniosków płynących z opracowania danych doświadczalnych. Czytelnik nie znajdzie tu zatem obszernych matematycznych wywodów poświęconych poszczególnym procedurom, ani też nie powinien oczekiwać, że opracowanie to jest pełnym kompendium wszystkich znanych metod analizy statystycznej. Dobór omawianych metod oraz przykładów ich wykorzystania w praktyce badawczej jest subiektywnym wyborem autora.

Z powyższych względów niezbędnym uzupełnieniem do tego opracowania powinny stać się znakomite obszerne podręczniki nowoczesnej statystyki, wymienione w wykazie piśmiennictwa. Obligatoryjnym uzupełnieniem są oczywiście tablice statystyczne, które są niezbędne przy rozwiązywaniu podanych przykładów i zadań analizy statystycznej.

Dla Czytelników zainteresowanych metodologią badań naukowych użytecznym uzupełnieniem „*Biostatystyki*" może być pozycja: Watała C., Różalski M., Boncler M., Kaźmierczak P. „*Badania i publikacje w naukach biomedycznych*" tom 1 i 2, α-medica press, 2011.

Dziękuję wszystkim, których sugestie i opinie, jak również rozmowy z nimi, skłoniły mnie do przygotowania tej książki.

prof. dr hab. Cezary Watała
Zakład Zaburzeń Krzepnięcia Krwi
Uniwersytet Medyczny w Łodzi

Łódź, grudzień 2011 r.

Wykaz stosowanych skrótów

χ^2	wartość statystyki chi^2
CI	przedział ufności
d.f. , *df*	liczba stopni swobody
F	statystyka testu *F* (Snedecora)
$f_{oczek(iwana)}$	liczebność oczekiwana
$f_{obs(erwowana)}$	liczebność obserwowana
f_i	liczba obserwacji w kategorii *i*
k	liczba klas/kategorii
μ	średnia populacji
MS	błąd średniokwadratowy
N	liczebność populacji (ogólnej, generalnej)
N_i, n_i	liczebność próby
n.s.	nieistotny (statystycznie)
p_i	proporcja obserwacji znalezionych w kategorii *i* (= f_i/n)
p, *q*	proporcje w próbie
R	suma rang
r_S	korelacja Spearmana
SD, *s*	odchylenie standardowe próby
SE	błąd standardowy
σ	odchylenie standardowe populacji
SS	suma kwadratów
t	statystyka testu *t* Studenta
x	wartość zmiennej
$\bar{X}_i$, $\overline{x_i}$	*i*-ta wartość zmiennej w próbie
$\bar{x}$, $\overline{X}$	średnia próby
z	statystyka testu normalnego

„O wielkości i pięknie nauki stanowi po części to, że poprzez własne krytyczne badanie dowiadujemy się, iż świat jest całkowicie odmienny od tego, co sobie kiedykolwiek wyobrażaliśmy – póki wyobraźni naszej nie poruszyło odrzucenie wcześniej przyjętych teorii."

Karl Raimund Popper

Część I

Podstawy teoretyczne metod statystycznych stosowanych w naukach biomedycznych

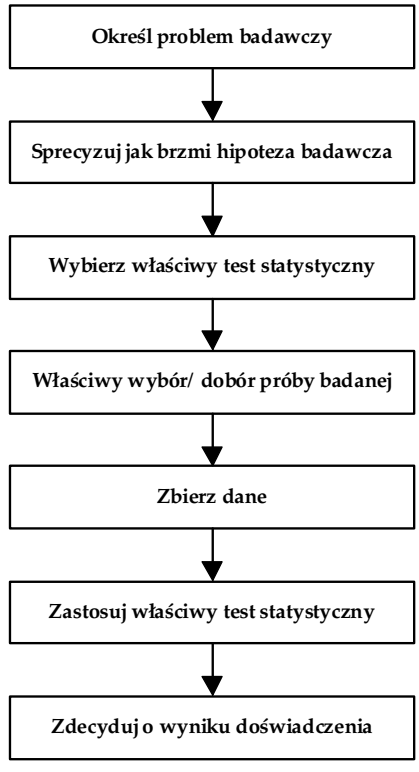
Rozdział 1

Podstawowe zasady planowania eksperymentu naukowego

„Musimy się uczyć z naszych eksperymentalnych błędów, które prowadzą do wyników niespełniających oczekiwań..."

Karl Raimund Popper

Schemat postępowania w planowaniu doświadczenia nie jest dowolny. Przystępując do wykonania eksperymentu naukowego w myśl sprawdzenia pomysłu, koncepcji czy hipotezy musimy wiedzieć, jak brzmi ta hipoteza i czego mamy bronić lub co obalać. Przypadkowe i szeroko zakrojone zbieranie wyników, a następnie porządkowanie ich w różnych konfiguracjach w celu sprawdzenia asocjacji czy różnic między nimi, jest nieporozumieniem i przeczy racjonalnemu podejściu w pracy naukowej. Wyniki doświadczeń naukowych zbieramy zawsze i opracowujemy z myślą o udowodnieniu postawionej wcześniej hipotezy, a nie odwrotnie. Próby *„wykrojenia"* koncepcji badawczej na podstawie zebranych wcześniej (najczęściej w sposób chaotyczny, gdyż bez wcześniejszego zamysłu celowego działania) danych pomiarowych jest działaniem po omacku i może sprawiać wrażenie manipulacji naukowej. W takim rozumieniu idei pracy naukowej przez doświadczonych badaczy, wdrukowanie sobie nawyku właściwego planowania pracy doświadczalnej jest podstawą uczciwej i wiarygodnej działalności naukowej. Oczywiście konieczność budowania schematu czy planu naukowego nie wyklucza sytuacji, gdy przypadkowo zgromadzimy dane doświadczalne pozwalające na udowodnienie jakiejś koncepcji, z której wcześniej nie zdawaliśmy sobie sprawy. Przykładem takiej sytuacji mogą być doniesienia kazuistyczne, których koncepcje powstają w oparciu o zbierane na gorąco obserwacje. Algorytm właściwego zaplanowania eksperymentu został zamieszczony na następnej stronie.



Określ problem badawczy

Co nas interesuje i czego jeszcze nie wiadomo? Co chcielibyśmy udowodnić? Niemożliwe jest uzyskanie nowej wiedzy na podstawie badań bez sprecyzowania tego, co chcemy zbadać – problem badawczy precyzujemy nie na podstawie tego co możemy zmierzyć mając do dyspozycji określony warsztat aparaturowy oraz zaplecze finansowe, lecz na podstawie tego co nas intryguje. A zatem musi nas w ogóle cokolwiek intrygować, abyśmy z przekonaniem oddali się działalności naukowej – musimy mieć coś, co nazywa się pasją badawczą. To niezwykle istotne, gdyż to właśnie ta pasja pcha nas w kierunku stawiania sobie kolejnych przyczynkowych zapytań w oparciu o gromadzone fakty – ona jest źródłem poczucia sensu tego co robimy – ona sprawia, że nie traktujemy nauki instytucjonalnie. Wiedza o problemie/przedmiocie badań oraz doświadczenie badawcze jest jedynie dopełnieniem – najczęściej przychodzą one z czasem – aczkolwiek to one decydują o udanym przedsięwzięciu naukowym – przychodzą tym szybciej im bardziej nasza pasja steruje naszym rozwojem naukowym. Można obrazowo powiedzieć, że cała otoczka instytucjonalna uprawiania nauki, ułatwiająca realizowanie zadań badawczych, jest jedynie dodatkiem do „*rdzenia*", którym jest kompetencja prowadzenia badań naukowych. Jak uczy doświadczenie, kompetencja taka przychodzi z czasem sama jako następstwo sukcesu naukowego. Wiadomo, że zależność ta jest jednokierunkową iteracją: kompetencja jest motorem polepszania warunków realizacji badań, ale nawet stworzenie najbardziej dogodnych warunków ekonomicznych nie jest samo w sobie w stanie doprowadzić do wykreowania doświadczonego pracownika naukowego.

Sprecyzuj, jak brzmi hipoteza badawcza

Co chcemy sprawdzić i udowodnić lub na jakie pytanie (pytania) odpowiedzieć?

Problemy ogólnej weryfikacji założeń hipotezy roboczej przedstawiono obszerniej w części *„Budowanie i weryfikacja hipotez badawczych"* (str. 56).

Wybierz właściwy test statystyczny

Czy model badawczy przystaje do modelu opracowania statystycznego wyników? Pytanie takie powinniśmy sobie postawić przystępując do każdego badania naukowego. Czy naprawdę jest konieczne zastanawianie się w jaki sposób będziemy analizować dane jeszcze przed ich zebraniem? Zdecydowanie tak, ponieważ decyzja o tym, czy będziemy mieli do czynienia ze zmiennymi ciągłymi czy dyskretnymi, zależnymi czy niezależnymi, itd. decyduje o tym jaki test można wykorzystać do analizy statystycznej.

Ocenia się, że około 60% wszystkich prac oryginalnych publikowanych w zakresie nauk biomedycznych czy farmakologicznych zawiera błędy opracowania statystycznego danych (De Muth, 1999). Najczęściej spotykane błędy w medycznej literaturze naukowej obejmują:

– niewłaściwe planowanie doświadczenia i/lub sformułowanie hipotezy badawczej,
– stosowanie błędu standardowego zamiast odchylenia standardowego lub odwrotnie jako miary rozproszenia,
– testowania hipotez statystycznych i doboru testów parametrycznych oraz nieparametrycznych,
– błędy wynikające z niespełnienia warunków normalności rozkładu i/lub jednorodności wariancji,
– niepoprawnego oszacowania lub nieoszacowania właściwej wielkość próby badanej,
– stosowania testów sparowanych i niesparowanych,
– stosowania wielokrotnych testów *t* jako rozwinięcia metod analizy wariancji,
– stosowania testu chi^2 i testu dokładnego Fishera.

Pomimo dużej różnorodności metod statystycznych niekiedy zdarza się, że układ doświadczalny nie spełnia idealnie warunków procedury statystycznej. Dotyczy to np. modeli analizy wariancji, porównań wielokrotnych, porównań wielu grup z powtórzeniami, itp. Niekiedy niepoprawny schemat wykonania doświadczenia może wręcz wykluczać poprawne użycie jakiegokolwiek dostępnego testu, a przynajmniej narzuca konieczność budowania indywidualnego modelu testu dostosowanego do analizy własnego układu eksperymentalnego. Pomijając fakt, że nie zawsze czujemy się kompetentni do wykonania takiej modyfikacji matematycznej, jest to niewygodne i niepraktyczne rozwiązanie – to jeszcze jeden argument, aby najpierw planować dobór metod statystycznych, a potem wykonywać doświadczenie.

Właściwe dobranie metody analizy statystycznej jest trudne i wymaga sporego doświadczenia i dobrej znajomości teoretycznych podstaw metod statystycznych stosowanych w badaniach naukowych. Przyjemną metodą pozyskiwania takiego doświadczenia, chociaż bardzo czasochłonną, może być przeglądanie przykładów opracowań statystycznych konkretnych wyników pracy naukowej w naukach medycznych. Aby przekazać Czytelnikowi niektóre wskazówki i praktyczne porady, w jaki sposób wybrane omawiane metody statystyczne możemy poprawnie stosować w praktyce badawczej i jak zrobić z nich najlepszy użytek, w drugiej części książki (*„Część II – Uzupełnienia, przykłady i zadania"*)

zamieszczono przykłady opracowania statystycznego konkretnych danych doświadczalnych.

Algorytmy doboru odpowiedniej metody analizy statystycznej zostały przedstawione w schematyczny sposób w części „*Algorytmy wyboru właściwego testu/metody analizy statystycznej*" (str. 32-35).

Właściwy wybór/dobór próby badanej

Należy sobie odpowiedzieć na pytania: Jakie cechy powinna mieć właściwie dobrana grupa kontrolna? Kto ma ją stanowić, aby była to reprezentatywna próba? Co ma decydować o tym, że tą grupę nazwiemy kontrolną – jakie kryteria? Czy to, że jej przedstawiciele nie mają określonej choroby, czy to, że nie biorą określonego leku?

Ile pomiarów musimy wykonać, aby udowodnić słuszność hipotezy statystycznej? Czy jeżeli przebadamy dużą (w naszym przekonaniu) liczbę osobników możemy mieć pewność, że wiarygodnie wypowiemy się o braku lub występowaniu różnic istotnych statystycznie? Na czym opieramy swoją ocenę, jak duża powinna być grupa? Czy prowadzimy badania dopóty, dopóki wystarczy nam środków finansowych, czy wykonujemy ściśle określoną liczbę pomiarów? Czy takiej estymacji dokonujemy *a priori*, czy *a posteriori*?

Do właściwej oceny pożądanej liczebności grupy służą metody estymacji liczebności próby badanej. Metody te zostały bliżej scharakteryzowane w Rozdziale „*Metody estymacji liczebności próby*". Stosujemy je m.in. po to, aby nie zbierać niepotrzebnie dużej liczby danych w sytuacji, gdy: 1) już na „pierwszy rzut oka" dostrzegamy, że porównywane grupy różnią się między sobą lub gdy 2) nie występują rzeczywiste różnice i nie wykażemy ich niezależnie od liczebności próby, a zbierając bardzo dużą liczbę powtórzeń mnożymy tylko niepotrzebnie koszty eksperymentu, podczas gdy moglibyśmy wykorzystać te środki na sprawdzenie innej koncepcji badawczej. Stosowanie estymacji właściwej liczebności próby powinno być nawykiem każdego rzetelnego badacza, a niewykorzystywanie tej metody może być uważane za niekompetencję w prowadzeniu badań naukowych.

Zbierz dane

Zasadą przeprowadzania badań klinicznych i biomedycznych jest często wykonanie oznaczeń wybranych parametrów w grupie pacjentów oraz w odpowiednio dobranej grupie kontrolnej, a także porównanie, czy badana jednostka chorobowa wpływa na zmiany tych parametrów. W analizie takiej powinniśmy uwzględnić dwa podstawowe wymagania opracowania statystycznego i porównania wyników badań:

- Ochotnicy do badań powinni być dobrani w sposób losowy.
  Planując przeprowadzenie badania należy odpowiedzieć sobie na pytania: Czy to w ogóle możliwe, aby dobrać ochotników do grupy kontrolnej w sposób całkowicie losowy? Jak dobrać grupy równocenne pod względem np. wieku, gdy badana jednostka chorobowa pojawia się jedynie w określonych grupach wiekowych? Czy w porównywalnej pod względem wieku grupie kontrolnej możemy wtedy wykluczyć wystąpienie jakichkolwiek czynników, które mogłyby wpływać na badane przez nas parametry? Jak przeprowadzić randomizację, jeżeli pacjenci, którzy reprezentują interesującą nas grupę badaną spotykani są jedynie sporadycznie w materiale klinicznym? Jak przeprowadzać właściwie testowanie wpływu leku (rola *placebo*)? Co to jest pojedyncza i podwójna ślepa próba?

- Każdy pomiar jest niezależny od innych pomiarów, chyba że mamy do czynienia z danymi sparowanymi lub próbami z powtórzeniami; niezależność danych to podstawowe założenie większości testów statystycznych – jeżeli jest niespełnione, to ryzykujemy, że wynik wnioskowania statystycznego będzie obarczony błędem.

Zastosuj właściwy test statystyczny

O poprawnym wyborze testu czy metody statystycznej analizy danych decyduje wiele czynników i często niewłaściwie przeprowadzone doświadczenie – tzn. wg niewłaściwego schematu badania różnic czy zależności – może dyskwalifikować wnioski badania naukowego.

Zadecyduj o wyniku doświadczenia

Weryfikacja poszczególnych poprawnie postawionych cząstkowych hipotez statystycznych pozwala opisać model, ponieważ umożliwia zweryfikowanie hipotezy opisującej zachodzenie procesu, prawidłowości, zależności, itp. Pamiętajmy, że możliwe jest jedynie odrzucenie hipotezy zerowej (z określonym prawdopodobieństwem), ale nigdy udowodnienie jej prawdziwości. Dlatego hipotezy statystyczne należy budować w ten sposób, aby ilość informacji wynikająca z ich odrzucenia była jak największa. Bardziej szczegółowe omówienie tego problemu znajdzie Czytelnik w części „*Budowanie i weryfikacja hipotez badawczych*" (str. 56).

Rodzaje badań naukowych oraz planowanie i prowadzenie badań w naukach biomedycznych

Wstępem do planowania doświadczenia czy badania naukowego powinno być sprecyzowanie celu badań. Co chcemy osiągnąć i na jakie pytania odpowiedzieć? Właściwe określenie problemu, jaki pragniemy rozwiązać prowadząc określone badania naukowe jest kluczem do sukcesu w znajdowaniu odpowiedzi na postawione sobie pytania. Celem prowadzenia badań w naukach biomedycznych może być:

1. charakterystyka określonej grupy badanej (populacji) i ustalenie interesujących nas cech tej grupy,
2. zbadanie związku między interesującymi nas parametrami oraz przeprowadzone na tej podstawie przewidywanie (predykcja) wartości interesujących nas zmiennych (parametrów) w przyszłości,
3. zbadanie skuteczności określonej strategii działania klinicznego, terapeutycznego, farmakologicznego, itp.

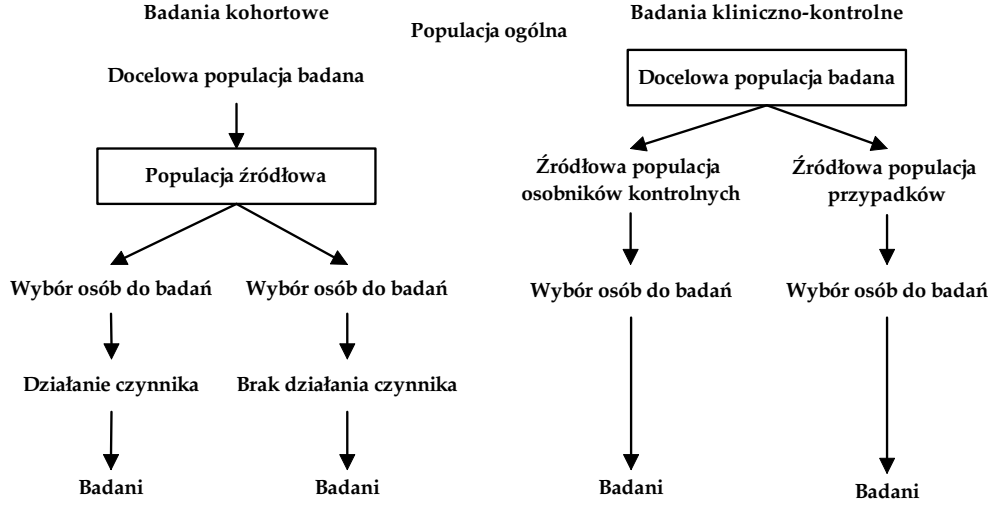
Badania naukowe w zakresie nauk biomedycznych możemy ogólnie podzielić na:

a) badania podstawowe (*basic research*) – takie, które przeprowadza się najczęściej w oparciu o model badawczy, czyli sztucznie stworzony układ „imitujący" lub symulujący warunki w ustroju; ten typ badań nakierowany jest z reguły na badanie molekularnych mechanizmów jakiegoś procesu; jego podstawową zaletą jest pełna manipulowalność parametrami układu, podstawową wadą natomiast jest niezadowalająca aproksymowalność układu modelowego do warunków naturalnych; taka niewystarczająca korespondencja między

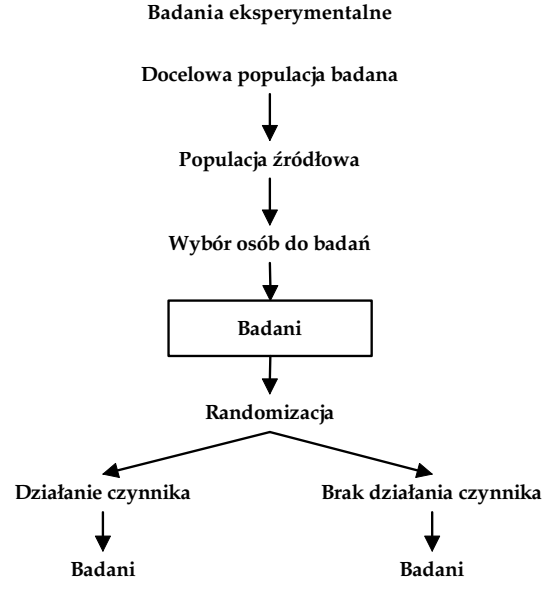
modelem a rzeczywistością prowadzi do ograniczonego zaufania do badań modelowych jako sposobu opisu działania układów biologicznych; z uwagi na swoją specyfikę są to bardziej badania typu przyczynkarskiego, a mniej całościowego;

b) badania populacyjne (*population studies*) i kliniczne (*clinical studies*) – pomimo bardzo licznych ograniczeń i problemów interpretacyjnych dotyczą one żywych obiektów biologicznych, stąd zaufanie do wyników tych badań ze strony badaczy-pragmatyków jest większe; badania tego typu mają charakter obserwacyjny, z minimalną, a przynajmniej ograniczoną ingerencją w układ badawczy; ich podstawową zaletą jest ścisła korespondencja z realnym układem biologicznym, wadą – niemożność modulowania zmian interesujących nas obserwowanych parametrów; tego typu badania są prawdziwym wyzwaniem dla tych badaczy, którzy poszukują ogólnej tendencji czy trendu (typowych dla dużych zbiorowości obiektów biologicznych, czyli populacji ogólnych); badania tego typu angażują złożony aparat statystyczny, którego poprawne zastosowanie i wszechstronne posłużenie się nim jest nie mniej ważne od samego gromadzenia wyników obserwacji i często wyrokuje o wartościach poznawczych całego badania.

Kryterium wydzielania różnych rodzajów badań populacyjnych jest poziom wybierania (selekcji) badanych obiektów. W celu przeprowadzenia badania typu populacyjnego z populacji ogólnej (generalnej) wydzielamy populację docelową, do której mają się odnosić wyniki badania. Jest to nasza reprezentatywna populacja, z której wylosujemy populację źródłową o zdefiniowanej liczebności i strukturze; z niej następnie wylosujemy badane osoby. Osoby wylosowane do badań w dalszym ciągu poddajemy selekcji, tak aby populacja uczestnicząca w badaniach została dobrana według ściśle zdefiniowanych kryteriów. Najprostszym sposobem charakterystyki badanej populacji może być analiza retrospektywna (*retrospective study*) lub prospektywna (*prospective study*) interesujących parametrów (zmiennych) tej populacji. Badania tego typu są z reguły mało kosztowne, ale ich podstawową wadą jest niemożność manipulowania modelem doświadczalnym przez badacza. Do takich badań zaliczamy badania kohortowe (*cohort studies*), w których wyboru przypadków dokonujemy na poziomie populacji źródłowej, oraz badania kliniczno-kontrolne (*case-control studies*), gdzie wyboru populacji źródłowej przypadków i populacji źródłowej kontroli dokonujemy na poziomie docelowej populacji badanej (*zobacz także schemat poniżej*).



Oba te rodzaje badań nazywamy badaniami typu obserwacyjnego (*observational studies*) (obserwowanie wyselekcjonowanych do badań przypadków bez interwencji czy manipulacji zmiennymi) i odróżniamy je od badań typu doświadczalnego (badania eksperymentalne, czyli takie, w których badacz manipuluje obiektem badanym i/lub zmienia wpływające nań czynniki).



W grupie badań obserwacyjnych możemy wyróżnić trzy podstawowe typy badań:

a) przekrojowe (*cross-sectional studies*), prowadzone przez krótki okres czasu i rejestrujące jakiś mały wycinek charakterystyki badanej populacji;
b) monitoringowe (*longitudinal studies, continuous monitoring studies*) (retrospektywne lub prospektywne), polegające na rejestrowaniu wybranych parametrów badanej populacji w dłuższym okresie czasu, często sprowadzają się do powtarzalnych badań typu przekrojowego; mogą być prowadzone zgodnie z modelem dynamicznym, kiedy skład badanej populacji zmienia się w czasie w myśl warunku spełniania przez objęte badaniami przypadki kryteriów badania, lub w oparciu o stałą nie uzupełnianą populację (kohortę); są bardziej złożonym podejściem badawczym, niż badania przekrojowe, statystyczna analiza wyników jest bardziej skomplikowana, wymagają znacznie dłuższego okresu prowadzenia badań, a więc są to badania bardziej kosztowne;
c) badania kliniczno-kontrolne (*case-control studies*), do analizy związku między występowaniem określonego czynnika a występowaniem choroby, obejmują przypadki (chorobowe, *cases*) i osobniki kontrolne; tego typu schematy badawcze sprawdzają się szczególnie w badaniach rzadkich chorób, ponieważ wymagają stosunkowo mało licznych prób w porównaniu z badaniami przekrojowymi lub monitoringowymi; kontrowersyjne jest zawsze dobranie właściwej grupy kontrolnej w tego rodzaju badaniach, badania te sprawdzają się najlepiej w wykrywaniu silnych i wyraźnych efektów.

Badania typu doświadczalnego (*experimental studies*) obejmują:

- wieloośrodkowe badania kliniczne (*clinical trials*),
- wieloośrodkowe badania dotyczące szczepionek (*vaccine trials*),
- badania interwencyjne i profilaktyczne (*intervention /prophylactic trials*).

Wieloośrodkowe badania kliniczne

Wieloośrodkowe badania kliniczne to eksperymenty, w których badamy skuteczność określonego (najczęściej nowego) sposobu leczenia: preparatu farmakologicznego, rodzaju terapii, dawki preparatu, itp. Ochotnicy, którzy wyrażają zgodę na wzięcie udziału w takich badaniach są w sposób losowy przydzielani do grupy kontrolnej (referencyjnej, w której stosujemy standardowe procedury, preparaty, dawki itp.) lub badanej (w której stosujemy zmienione/nowe procedury, preparaty, dawki itp.). W doświadczeniach takiego typu niezwykle istotne jest ustawienie schematu badań w taki sposób, aby interesujący nas efekt (wpływ zmienionej terapii, leku, itp.) można było przypisać działaniu tego właśnie czynnika, nie zaś różnicom charakterystyki włączonych do badań grup. Obie grupy powinny być jak najbardziej podobne do siebie pod względem takich cech, jak struktura wiekowa, rozkład płci, wskaźniki antropomorfometryczne czy nawet status genetyczny. Taki staranny dobór służy eliminacji określonych skłonności (tendencji, *bias*) w kierunku *nielosowości* doboru próby, która mogłaby zafałszować obserwowane prawidłowości. Poza tym, że przydział ochotników do grup powinien podlegać randomizacji, w badaniach dotyczących np. testowania wpływu określonego preparatu farmakologicznego, powinniśmy eliminować taką skłonność do nielosowości stosując sprawdzone algorytmy badania z ***pojedynczą*** lub ***podwójną ślepą próbą***. Na czym one polegają i jaka jest między nimi różnica? Aby wyeliminować wpływ czynnika psychologicznego w badaniach testowania wpływu preparatów farmakologicznych (szczególnie w sytuacjach, gdy istotną rolę w powodzeniu terapii czy leczenia odgrywają czynniki subiektywne), grupa kontrolna powinna być – dla osoby niewtajemniczonej – nieodróżnialna od grupy badanej. Dokonuje się tego w ten sposób, że osobom z grupy badanej podaje się testowany lek, zaś osobom z grupy kontrolnej *placebo.* Schemat ***podwójnej ślepej próby*** (*double-blind design*) przewiduje, aby ani osoba badana, ani badacz nie wiedzieli kto otrzymuje lek, a kto *placebo*, aż do czasu zakończenia doświadczenia. W sytuacji gdy jedynie osoba badana nie ma świadomości jaki preparat przyjmuje, mówimy o ***pojedynczej ślepej próbie*** (*single-blind design*).

Randomizację doboru ochotników (*random sampling*) oraz ich zaszeregowania do jednej z grup mogą nam ułatwić tablice liczb losowych albo komputerowe generatory liczb losowych. Sposób wykorzystywania tych „wygenerowanych" lub wyszukanych liczb losowych pozostawia dużo dowolności: możemy na przykład przydzielać numery parzyste do grupy kontrolnej, a nieparzyste do grupy badanej. Jeżeli na przykład, wyszukane liczby losowe tworzą ciąg cyfr 4, 7, 0, 9, 5, 2, to pierwszego, trzeciego i ostatniego ochotnika przydzielilibyśmy do grupy kontrolnej, zaś pozostałych (drugiego, czwartego i piątego) do grupy badanej. Przy takim postępowaniu możemy co jakiś czas – na przykład co 10 wylosowanych numerów – zastosować restrukturyzację naszych losowych przydziałów, aby mieć pewność, że do każdej z grup trafia tyle samo ochotników, lub by mieć pewność, że proporcje kobiet i mężczyzn są podobne. Jest to tak zwana ograniczona randomizacja lub randomizacja zrównoważona (*restricted randomization, randomization with balance*), gdyż ingerujemy w rzeczywisty losowy dobór ochotników do każdej z grup. W takim przypadku

korzystanie z tablic liczb losowych jest trochę inne: dla każdej 10-tki losujemy 5 liczb i przyporządkowujemy je np. ochotnikom z grupy badanej. Powiedzmy, że będą to liczby 08, 01, 03, 06, 04, czyli ósmy, pierwszy, trzeci, szósty i czwarty ochotnik trafia do grupy badanej, zaś pozostałe, czyli drugi, piąty, siódmy, dziewiąty i dziesiąty trafią do grupy kontrolnej. Taką „randomizację dziesiątkami" powtórzymy dla każdych dziesięciu ochotników.

Aby zapewnić równe proporcje kobiet i mężczyzn możemy randomizację według powyższego schematu przeprowadzić osobno dla kobiet i dla mężczyzn.

Zazwyczaj staramy się dobrać grupę kontrolną i badaną podobnie pod względem proporcji płci oraz struktury wiekowej, głównie dlatego, aby w miarę możliwości wyeliminować wpływ wszystkich innych czynników dodatkowych poza tym, którego efekt badamy. Takie sparowanie pod względem płci i wieku (*matched-pair design*) jest zresztą ogólnym wymaganiem w badaniach populacji żywych organizmów, także w badaniach porównawczych typu obserwacyjnego staramy się tego wymagania nie naruszać.

Jeżeli nasz eksperyment dotyczący wpływu jakiegoś czynnika nie trwa bardzo długo, bardzo pożądanym jest zastosowanie modelu krzyżowego (*cross-over design*), w którym ochotnicy badani stanowią swoja własną kontrolę. W takich przypadkach istotne jest przestrzeganie określonego okresu „spoczynkowego" (*washout*), aby efekty dwóch porównywanych czynników nie nałożyły się.

Nie zawsze badania tego typu obejmują zdrowych ochotników, z których część otrzymuje testowany preparat, a inni *placebo*. Niekiedy schemat badania zakłada, że badany preparat farmakologiczny otrzymują chorzy pacjenci oraz zdrowi ochotnicy. Ten wariant modelu badawczego wiąże się z problemami natury etycznej, szczególnie w stosunku do zdrowych ochotników: jak długo prowadzić eksperyment skoro z jednej strony wiadomo, że jego wyniki mogą przynieść wymierne korzystne efekty dla pacjentów, z drugiej zaś poddawanie zdrowych ochotników długotrwałemu reżimowi terapeutycznemu może nie być zupełnie obojętne?

Rozdział 2

Podstawowe pojęcia i terminy stosowane w statystyce

Zmienne i ich rodzaje

Zmienne to wielkości (parametry, cechy), które mierzymy, kontrolujemy lub którymi manipulujemy w jakiś sposób w trakcie badań. Zależnie od tego, jakiego parametru dotyczą, zmienne mogą mieć różne właściwości, zarówno ze względu na rolę jaką pełnią w naszych badaniach, jak też ze względu na to, jaki rodzaj miary można do nich zastosować.

Ogólnie zmienne zaliczamy do jednej z dwóch kategorii:
1. zmienne zależne (*dependent variable*),
2. zmienne niezależne (*independent variable*).

Rozróżnienie między nimi wydaje się terminologicznie dość paradoksalne, gdyż wartości każdej zmiennej zależą od wartości innych zmiennych. Określenia „*zależny*" i „*niezależny*" mają zastosowanie głównie w badaniach typu eksperymentalnego. Niezależnymi nazywamy takie zmienne, których wartości możemy dobierać i zmieniać w doświadczeniu (są to zmienne manipulowane przez badacza). Zmienne zależne natomiast mogą być jedynie mierzone lub rejestrowane przez badacza, nie ma on wpływu na to, jakie wartości przyjmują. Zmienne, które może zmieniać badacz są niezależne od początkowych cech, wzorców, skłonności, itp. badanych obiektów. Inne zmienne będą zależne od warunków i okoliczności eksperymentu. Możemy powiedzieć, że zależą one od tego, co obiekt uczyni lub jak się zachowa w odpowiedzi na zadane przez badacza zmiany. Dość niefortunnie – i niejako w opozycji do tego rozróżnienia – terminy te bywają również stosowane w badaniach, gdzie nie manipuluje się dosłownie zmiennymi niezależnymi, lecz jedynie przypisuje się obiekty do pewnych grup doświadczalnych, np. badając wartości hematokrytu u kobiet i mężczyzn, płeć można traktować jako zmienną niezależną, zaś hematokryt jako zmienną zależną.

Zmienne różnią się także pod względem ilości mierzalnej informacji, którą można zdobyć w następstwie ich pomiaru. Pod tym względem zmienne można podzielić na: 1) nominalne, 2) porządkowe, 3) przedziałowe, 4) ilorazowe. Pierwsze dwa rodzaje reprezentują ***zmienne dyskretne*** *(discrete variables)*, czyli takie, które zmieniają się skokowo i mogą przyjmować jedynie określone wartości zbioru liczb całkowitych. Zmienne przedziałowe oraz ilorazowe to ***zmienne ciągłe*** *(continuous variables)*, przyjmujące dowolne wartości z określonego przedziału zbioru liczb rzeczywistych.

1. Zmienne nominalne (*data on nominal scale*) mogą być mierzone jedynie w kategoriach zaklasyfikowania poszczególnych egzemplarzy do jednej z rozróżnialnych kategorii, lecz już w ramach tej kategorii nie można ich klasyfikować ilościowo ani nawet nadać im rang. Przykładem takiej zmiennej może być zmienna (A) określająca typ antropoidalny; w jej kategoriach można zakwalifikować określonych osobników do rasy nordyckiej lub negroidalnej, ale nie możemy już powiedzieć, który z nich posiada własność niesioną przez zmienną A w większym stopniu. Typowymi przykładami zmiennych nominalnych są płeć, rasa, kolor włosów, miasto zamieszkania, itp.
2. Zmienne porządkowe (*data on ordinal scale*) pozwalają na rangowanie (ustawianie w określonym porządku) badanych elementów, w tym sensie, że element z wyższą rangą posiada cechę reprezentowaną przez mierzoną zmienną w większym stopniu, którego jednak nie wyrażamy w sposób ilościowy. Przykładem zmiennej porządkowej może być stopień pigmentacji skóry lub stopień otyłości.
3. Zmienne przedziałowe (*data on interval scale*) pozwalają nie tylko szeregować (rangować) mierzone elementy, lecz również mierzyć różnice wielkości pomiędzy nimi. Na przykład temperatura mierzona w skali Celsjusza jest zmienną przedziałową. Można powiedzieć, że wzrost temperatury od 20 do 40 stopni jest dwa razy tak duży jak od 30 do 40 stopni.
4. Zmienne ilorazowe (*data on ratio scale*) są podobne do zmiennych przedziałowych, lecz charakteryzuje je ponadto istnienie punktu absolutnego zera skali, dzięki czemu prawomocne jest w odniesieniu do tych zmiennych stwierdzenie typu: x jest dwa razy większe niż y. Typowymi przykładami skal ilorazowych są skale przestrzeni, czasu, czy temperatury mierzonej w skali Kelvina. Można powiedzieć, że 200 stopni Kelvina jest temperaturą dwa razy wyższą niż 100 stopni Kelvina. W większości procedur/testów statystycznych nie dokonuje się rozróżnienia pomiędzy skalą przedziałową i ilorazową.

Miary położenia i rozproszenia

Zwyczajowo charakteryzuje się (sub)populację/grupę wyników podając miarę położenia, wyrażającą tendencję centralną zmiennej, i towarzyszącą jej miarę rozproszenia, wyrażającą zmienność wewnątrzgrupową. Można także dodatkowo podawać zakres wartości (wartość minimalna-wartość maksymalna, dolny kwartyl, górny kwartyl).

Miary położenia *(measures of central tendency)*

Średnie, mediana i modalna

Najczęściej stosowaną miarą centralną jest ***średnia arytmetyczna*** (średnia – *arithmetic mean*):

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

gdzie Σx_i oznacza sumę poszczególnych wartości zmiennej, natomiast n – liczebność grupy/próby.

W przypadku obliczania wartości średniej na podstawie częstości (f_i) wyników w próbie stosujemy średnią ważoną.

$$\bar{x} = \frac{\sum_{i=1}^{k} f_i x_i}{n}$$

Kiedy rozkład wyników jest prawoskośny, lepszym odzwierciedleniem tendencji centralnej jest ***średnia geometryczna*** *(geometric mean)*:

$$\bar{x}_G = anti\log(\bar{u}) = 10^{\bar{u}}, \text{ gdzie } \bar{u} = \frac{\sum_{i=1}^{n} u_i}{n} \text{ oraz } u_i = \log(x_i).$$

Ten typ średniej często wykorzystuje się do analizy wyników tworzących ciągi geometryczne, np. miareczkowanie miana antygenu, przeciwciał, itp.

Średnia harmoniczna *(harmonic mean)* jest stosowana najczęściej do uśredniania częstości.

$$\bar{X}_H = \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}} = \frac{n}{\sum_{i=1}^{n}\frac{1}{x_i}}$$

Jeżeli dane są identyczne, to średnie arytmetyczna, geometryczna i harmoniczna są sobie równe; dla nieidentycznych danych $\bar{\bar{X}}_H < \bar{X}_G < \bar{X}$.

Mediana (*median*) jest środkową wartością w szeregu rosnącym (lub malejącym) wyników. Możemy zapisać, że

$Me = \frac{(n+1)}{2}$ – ta obserwacja uporządkowanego ciągu wyników lub $Me = X_{(n+1)/2}$

Jeżeli n jest nieparzyste, wówczas wyrażenie $\frac{(n+1)}{2}$ jest liczba całkowitą wyznaczającą położenie środkowej wartości w uporządkowanym szeregu wyników. Jeżeli natomiast n jest parzyste, wówczas wyrażenie $\frac{(n+1)}{2}$ nie jest liczba całkowitą i przypada między jedną a drugą wartością w uporządkowanym szeregu wyników. Medianę wyznacza się wówczas jako średnią pary wartości sąsiadujących.

Mediana jest szczególnym kwantylem dzielącym uporządkowaną zbiorowość wyników na dwie równe części.

Mediana jest użyteczna w przypadkach, gdy występują pojedyncze skrajnie niskie lub wysokie wartości, niereprezentatywne dla całej próby, zaniżające lub zawyżające wartość średniej.

Modalna *(mode)* jest najczęściej występującą wartością w grupie.

Średnia arytmetyczna, mediana i modalna są w przybliżeniu równe, kiedy rozkład wyników jest symetryczny i jednomodalny.

Kwantyle (*quantiles*)

Podobnie jak w przypadku mediany, możemy definiować wartości poniżej i powyżej których leżą frakcje wyników podzielonych na kilka równych części. I tak, dzieląc zbiorowość danych na 4, 5, 10 czy 100 takich równych części mówimy o kwartylach (*quartiles*), kwintylach, decylach lub percentylach. Na przykład, jeżeli podzielimy uporządkowane dane na cztery równe części, to jedna czwarta tych wyników będzie miała wartości mniejsze niż pierwszy kwartyl (Q_1), jedna czwarta leży między Q_2 i Q_3, jedna czwarta między Q_3 i Q_4, zaś reszta powyżej Q_4. Pierwszy kwartyl możemy zapisać jako:

$$Q_1 = X_{(n+1)/4}.$$

Obliczona wartość indeksu dolnego $(n+1)/4$ jest zaokrąglana do najbliższej liczby całkowitej lub półcałkowitej.

Z powyższego wynika, że wartość Q_2 jest identyczna jak wartość mediany. Wartość LD_{50}, często stosowana w naukach przyrodniczych, jest 50-tym percentylem dawek letalnych lub środkową dawką letalną, co oznacza, że 50% testowanych obiektów przeżyło tę dawkę, a pozostałe 50% nie.

Miary rozproszenia (*measures of dispersion and variability*)

Wariancja i odchylenie standardowe

Wariancja *(variance)* określa rozrzut wyników wokół wartości średniej:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{(n-1)}.$$

Wyrażenie mianownika (*n–1*) zwane jest ***liczbą stopni swobody*** (*degrees of freedom*) wariancji i oznacza, że jedynie (*n–1*) z całkowitej liczby n odchyleń $(x_i - \bar{x})$ jest wzajemnie niezależnych od siebie, to znaczy takich, co do których możemy dokonać całkowicie wolnych (swobodnych) wyborów między jednym a drugim. Ostatnie wyrażenie $(x_i - \bar{x})$ nie może już być swobodnie wybrane spomiędzy innych, ponieważ jest jedyne; może ono być obliczone na podstawie innych, przy założeniu, że suma wszystkich n odchyleń wynosi zero.

Jednorodność wariancji (*homoscedastyczność, homogeneity of variance, homoscedasticity*) oznacza, że wariancje obrazujące zmienność wyników w kilku różnych (sub)populacjach/grupach nie różnią się istotnie. Występowanie takich różnic wariancji nazywamy *heteroscedastycznością* (*heterogeneity of variance, heteroscedasticity*).

Odchylenie standardowe *(standard deviation, SD)*

Jest to najpowszechniej stosowana miara zmienności wewnątrzgrupowej.

$$SD, s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{(n-1)}}$$

$$s = \sqrt{\frac{\sum x^2 - (\sum x)^2 / n}{(n-1)}}$$

W praktyce odchylenie standardowe interpretuje się jako miarę rozrzutu, która wyznacza przedział wokół średniej obejmujący około 70% wszystkich obserwacji (przy spełnieniu warunków normalności rozkładu). Zakres średniej ± 2 SD obejmuje 95% obserwacji i nazywany jest ***normą*** (*normal range*).

Współczynnik zmienności *(coefficient of variation, CV)*

Mówi o tym jaki procent średniej stanowi zmienność próby wyrażaną jako odchylenie standardowe.

$$CV = \frac{s}{x} \times 100\,\%$$

Użyteczność tej miary może zobrazować przykład dotyczący porównania zmienności wielkości małżowin usznych u ryjówek i słoni afrykańskich. Bezwzględne wartości SD u słoni są kilkaset razy większe, zaś wariancje nawet kilkanaście tysięcy razy większe, nie oznacza to jednak, że zmienność tej cechy jest minimalna u ryjówek i ogromna u słoni.

W niektórych dziedzinach, np. w badaniach cytomorfometrycznych, miara ta jest nawet chętniej i częściej stosowana niż odchylenie standardowe.

Wskaźniki (***indeksy***) ***różnorodności*** *(indices of diversity)*

Cechy o rozkładzie równomiernym obserwacji wśród wielu klas/kategorii charakteryzują się wysokimi wskaźnikami różnorodności. Te, dla których olbrzymia większość obserwacji przypada na kilka (nielicznych) klas/kategorii posiadają niską różnorodność. Najpowszechniej stosowanym wskaźnikiem różnorodności jest wskaźnik Shannona-Wienera:

$$H' = -\sum_{i=1}^{k} p_i \log p_i$$

lub

$$H' = \frac{n \log n - \sum_{i=1}^{k} f_i \log f_i}{n}$$

gdzie:

k oznacza liczbę klas/kategorii,

n liczebność próby,

f_i jest liczbą obserwacji w kategorii i, zaś

$p_i = f_i/n$ jest proporcją obserwacji znalezionych w kategorii i.

Na wielkość H′ ma wpływ nie tylko rozkład danych, ale także liczba kategorii; stąd maksymalną różnorodność określa równanie:

$$H'_{\max} = \log k\,.$$

Indeks $J' = \frac{H'}{H'_{\max}}$ nazywamy równomiernością lub homogennością rozkładu (*homogeneity, relative diversity*), zaś *1–J'* jest miarą heterogenności lub dominacji (*dominance, heterogeneity*). Ogólnie, ponieważ liczba klas/kategorii *k* jest z reguły niedoszacowana dla populacji, *J'* jest najczęściej przeszacowanym wskaźnikiem równomierności w populacji.

Dla prób z obserwacjami uzyskanymi w sposób nielosowy stosuje się wskaźnik Brillouina:

$$H = \frac{\log\left(\frac{n!}{\prod_{i=1}^{k} f_i!}\right)}{n}$$

lub

$$H = \frac{\log \frac{n!}{f_1! f_2! \ldots f_k!}}{n}$$

lub

$$H = \frac{(\log n! - \sum \log f_i!)}{n}.$$

Więcej informacji na temat wskaźników różnorodności może Czytelnik znaleźć w podręcznikach Zara oraz Sokala i Rohlfa.

Błąd standardowy i precyzja określania wartości średniej próby

Zarówno szacowana wartość średnia, jak i SD są jedynie z pewnym przybliżeniem reprezentatywne dla populacji ogólnej, a czym mniejsza jest badana próba, w której takiej estymacji dokonano, tym większe prawdopodobieństwo, że estymowane wartości $\bar{x}$ oraz SD będą się różniły od wartości rzeczywistych m oraz s. Jest bardzo prawdopodobne, że w innej próbie losowej szacowane wartości $\bar{x}$ oraz SD będą inne, a różnice takie wynikają ze zmienności zbierania obserwacji w poszczególnych próbach losowych. Jeżeli przeprowadzimy wiele losowań próby z populacji ogólnej, i w każdej tak wylosowanej próbie określimy wartość $\bar{x}$, to okaże się, że średnia obliczona na podstawie rozkładu częstości tych cząstkowych średnich będzie bliska średniej populacji, m. Odchylenie standardowe takiego rozkładu, zwane ***błędem standardowym średniej*** (*standard error, SE; standard error of mean, SEM*) będzie wynosić $\sigma / \sqrt{n}$. SE charakteryzuje błąd, z jakim szacujemy średnią dla populacji ogólnej, tzn. określa rozrzut między wieloma średnimi określanymi dla wielu małych grup losowanych z tej populacji. Zatem SE jest miarą tego, jak dokładnie szacujemy miarę centralną populacji ogólnej na podstawie określania średniej dla próby losowej. SE zależy zarówno od ogólnej zmienności wewnątrzpopulacyjnej, jak i od wielkości próby losowej.

W przypadku, gdy badane grupy są bardzo mało liczne, a w związku z tym istnieje duże prawdopodobieństwo, że średnie i SD estymowane dla tak mało licznych grup są słabo reprezentatywne dla całej populacji, podawanie SE ma większy sens statystyczny niż podawanie SD.

Jeżeli badamy próbę losowaną z populacji o skończonej wielkości, zmienność losowania próby jest znacząco niższa niż $\sigma / \sqrt{n}$, ponieważ n stanowi wysoką proporcję całkowitej liczebności populacji, N. Zmienność taka będzie wynosiła zero, jeżeli przebadamy całą taką skończoną populację, nie dlatego, że zmienność cechy wewnątrz populacji nie występuje, ale dlatego, że średnia dla próby jest *de facto* równa średniej dla populacji.

W sytuacjach takich zasadne jest stosować korekcję wartości SE dla skończonej populacji:

$$SE_{kor} = \frac{\sigma}{\sqrt{n}} \sqrt{\left(1 - \frac{n}{N}\right)}.$$

Jak widać w Tabeli poniżej, stosowanie poprawki ma niewielki wpływ w próbach losowych stanowiących mniej niż 10% ogólnej populacji i może być wtedy zaniedbane.

SE bez poprawki	**SE z poprawką**	**frakcja próby losowej**
1	0.99	1%
1	0.97	5%
1	0.95	10%
1	0.87	25%
1	0.71	50%
1	0.50	75%
1	0.32	90%
1	0.22	95%
1	0.10	99%
1	0.00	100%

Rozdział 3

Wybrane zagadnienia i problemy doboru właściwych metod statystycznych w opracowaniach wyników badań biomedycznych

Algorytmy wyboru właściwego testu/metody analizy statystycznej

Jak wynika z Rozdziału 1 („*Podstawowe zasady planowania eksperymentu naukowego*", str. 15), właściwe zaplanowanie wszystkich etapów pracy badawczej, włączając etap analizy danych doświadczalnych, jest warunkiem powodzenia w tej pracy, a także swoistym wskaźnikiem kompetencji w prowadzeniu badań naukowych.

Powszechnym błędem popełnianym w pracy naukowo-badawczej jest zbieranie danych najpierw – często w sposób dość przypadkowy – a dopiero później dopasowywanie testu, który możemy wykorzystać do ich opracowania. Ocena statystyczna nie może być dokonana dla danych zebranych w sposób arbitralny i przypadkowy, ponieważ każdy test statystyczny ma swoje wymagania, a ich niespełnianie dyskwalifikuje często wiarygodność tego testu.

Na przykład, wykonujemy pomiary wpływu kilku stężeń różnych leków na odpowiedź komórek. W zależności od tego, co zamierzamy udowodnić, grupą kontrolną mogą być:

a) komórki niepoddane działaniu żadnego leku (gdy sprawdzamy jak działają poszczególne leki),
b) komórki poddane działaniu określonego leku, o którym wiadomo jak działa – wtedy gdy chcemy wiedzieć, czy nowy badany związek działa podobnie czy inaczej,
c) komórki poddane działaniu leku w niskim stężeniu – wtedy gdy chcemy sprawdzić, czy zwiększenie stężenia leku zmienia efekt.

Czy badając wpływ poszczególnych leków powinniśmy dla każdego dobrać grupę kontrolną – wtedy gdy do opracowania stosujemy test sparowany – czy też powinna być jedna grupa kontrolna dla różnych stężeń różnych leków – wtedy gdy wyniki chcemy opracować metodą zrandomizowanej analizy blokowej? Czy pomiary będą wykonywane w kilku powtórzeniach czy w jednym – zależnie od tego jaki model analizy wariancji chcemy wykorzystać?

Częstym dysonansem utrudniającym dobór właściwej metody statystycznej niedoświadczonym badaczom może być także brak normalności rozkładu lub nierówne miary rozrzutu (wariancje) w poszczególnych grupach. W miarę nabywania doświadczenia, uczymy się jak radzić sobie z takimi „zakłóceniami", na przykład na drodze zwiększania liczebności próby lub poprzez transformację danych.

Właściwe dobranie metody lub grupy metod analizy statystycznej może ułatwić praktyczny schematyczny przewodnik wyboru metod analizy statystycznej, zamieszczony na stronach 32-35 tego opracowania (diagramy A, B, C i D). Należy sobie oczywiście zdawać sprawę, że schematy te to nie klucz do bezbłędnego wybierania odpowiedniego testu statystycznego – mają one raczej skłonić Czytelnika do bliższego zainteresowania się określoną grupą metod analizy statystycznej.

Zdaniem autora bardzo przydatne będzie również zajrzenie do drugiej części opracowania (*„Część II – Uzupełnienia, przykłady i zadania"*), gdzie zamieszczono przykłady wykorzystania poszczególnych metod statystycznych do analizy konkretnych danych doświadczalnych.

Diagram A

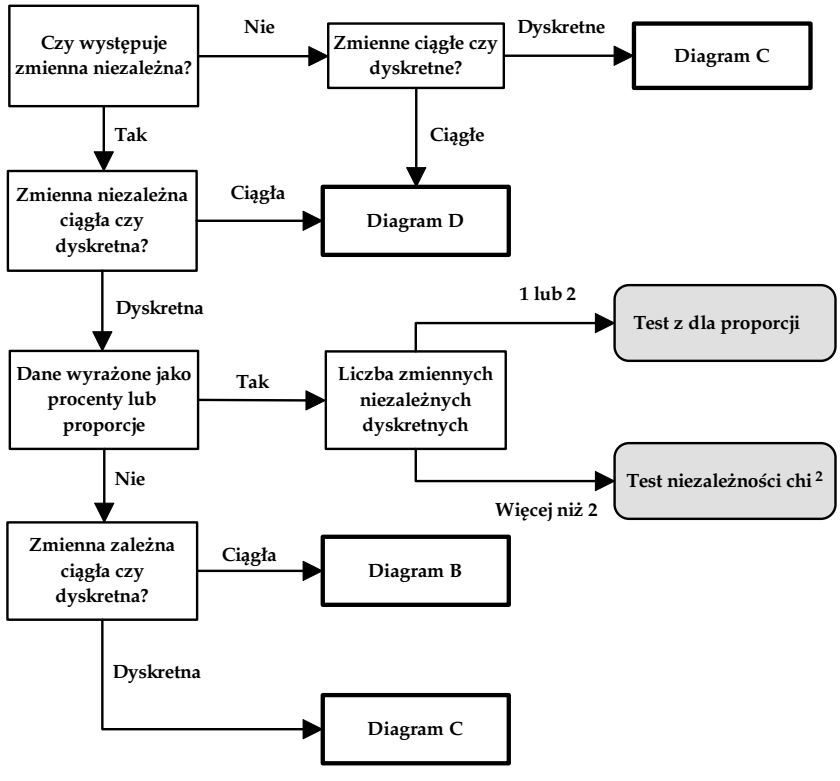


Diagram B

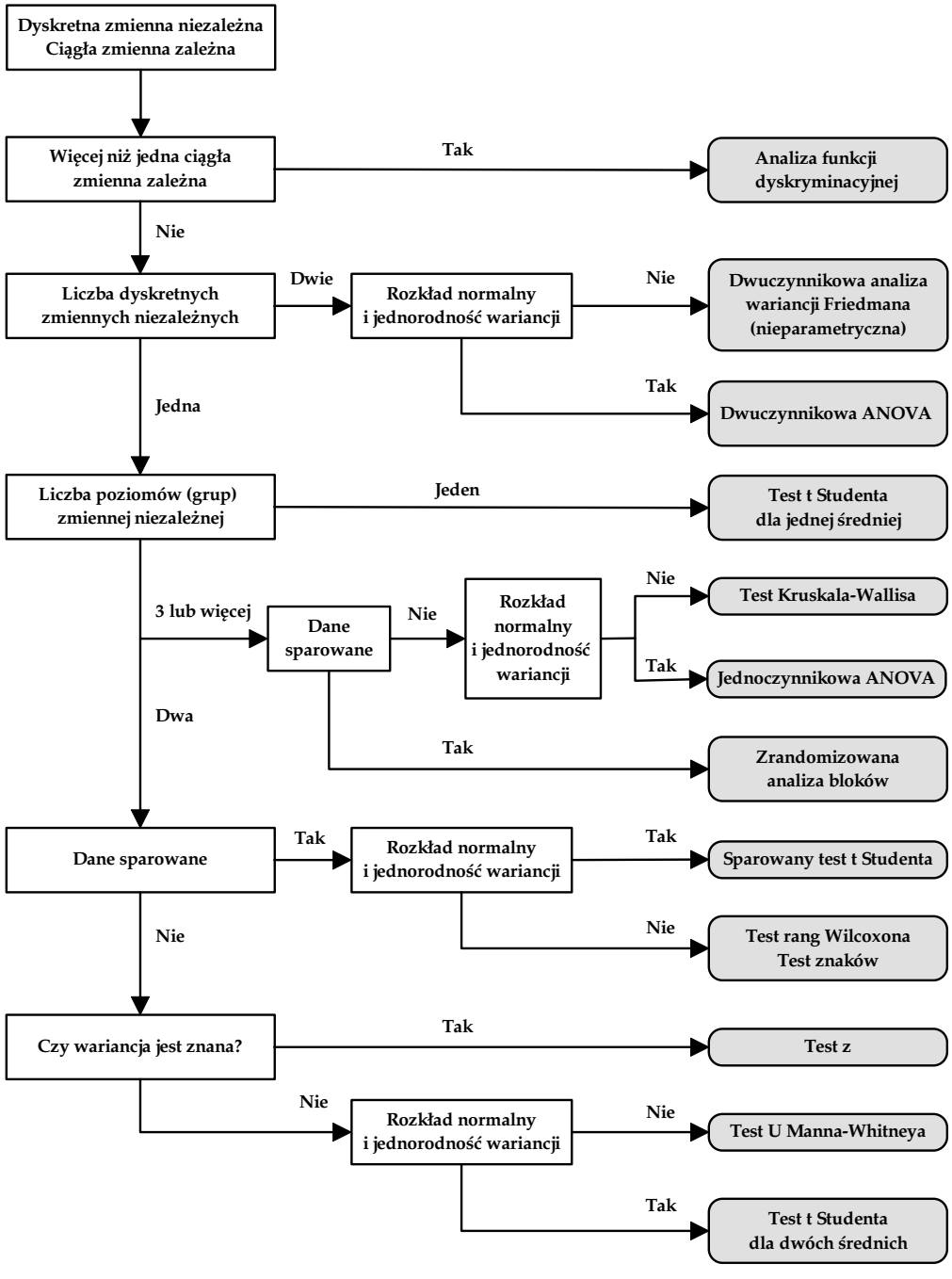

Dyskretna zmienna niezależna
Ciągła zmienna zależna
Więcej niż jedna ciągła zmienna zależna
Tak
Analiza funkcji dyskryminacyjnej
Nie
Liczba dyskretnych zmiennych niezależnych
Dwie
Rozkład normalny i jednorodność wariancji
Nie
Dwuczynnikowa analiza wariancji Friedmana (nieparametryczna)
Tak
Dwuczynnikowa ANOVA
Jedna
Liczba poziomów (grup) zmiennej niezależnej
Jeden
Test t Studenta dla jednej średniej
3 lub więcej
Dane sparowane
Nie
Rozkład normalny i jednorodność wariancji
Nie
Test Kruskala-Wallisa
Tak
Jednoczynnikowa ANOVA
Dwa
Tak
Zrandomizowana analiza bloków
Dane sparowane
Tak
Rozkład normalny i jednorodność wariancji
Tak
Sparowany test t Studenta
Nie
Nie
Test rang Wilcoxona
Test znaków
Czy wariancja jest znana?
Tak
Test z
Nie
Rozkład normalny i jednorodność wariancji
Nie
Test U Manna-Whitneya
Tak
Test t Studenta dla dwóch średnich

Diagram C

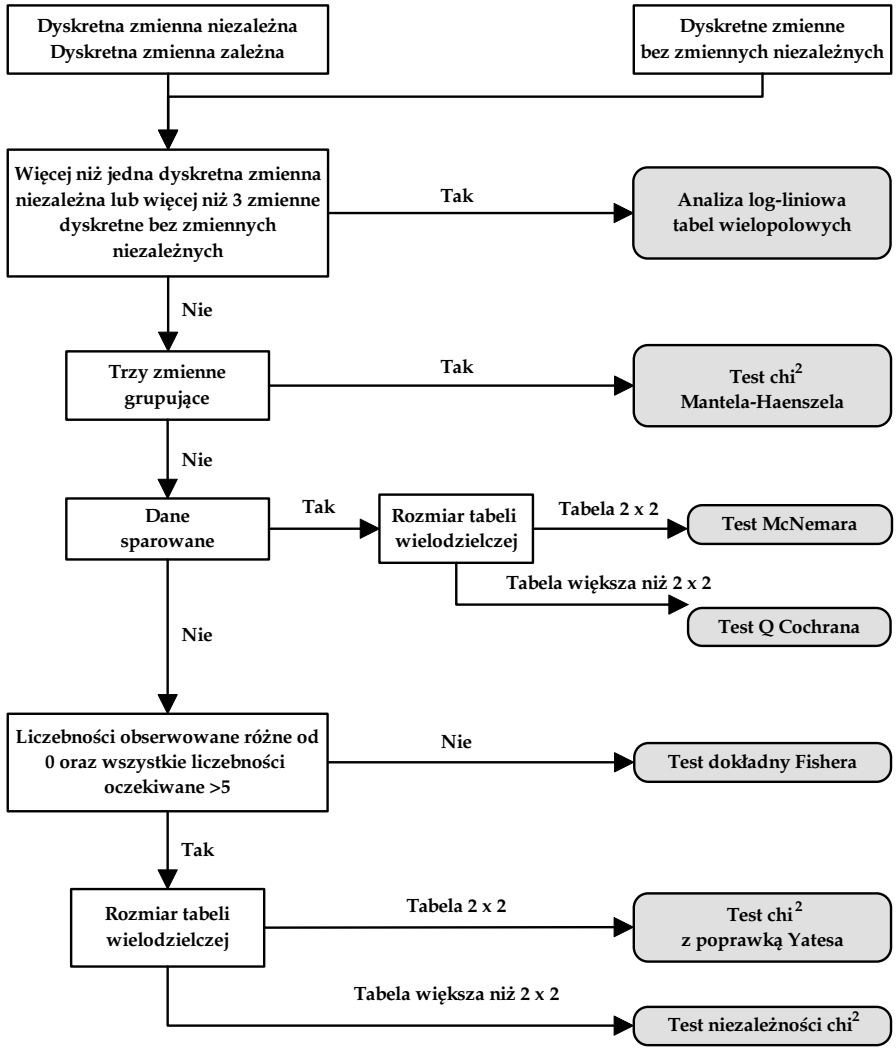
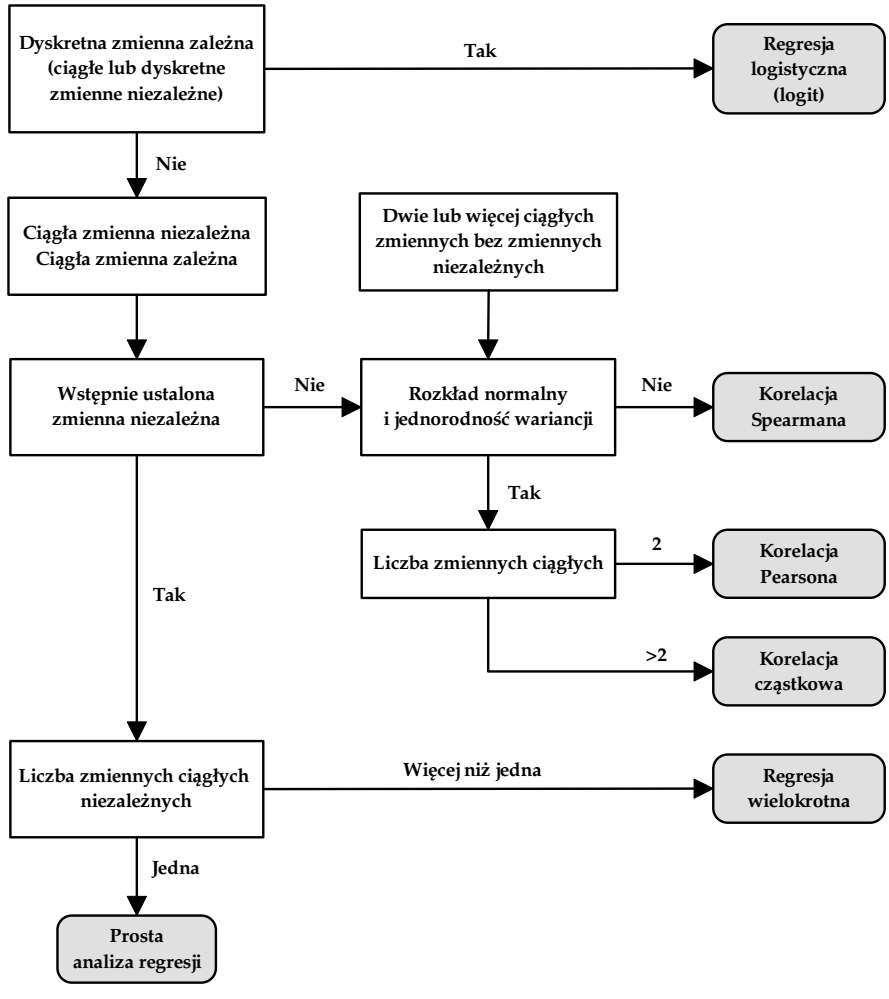

Dyskretna zmienna niezależna
Dyskretna zmienna zależna
Dyskretne zmienne
bez zmiennych niezależnych
Więcej niż jedna dyskretna zmienna niezależna lub więcej niż 3 zmienne dyskretne bez zmiennych niezależnych
Tak
Analiza log-liniowa tabel wielopolowych
Nie
Trzy zmienne grupujące
Tak
Test chi2 Mantela-Haenszela
Nie
Dane sparowane
Tak
Rozmiar tabeli wielodzielczej
Tabela 2 x 2
Test McNemara
Tabela większa niż 2 x 2
Test Q Cochrana
Nie
Liczebności obserwowane różne od 0 oraz wszystkie liczebności oczekiwane >5
Nie
Test dokładny Fishera
Tak
Rozmiar tabeli wielodzielczej
Tabela 2 x 2
Test chi2 z poprawką Yatesa
Tabela większa niż 2 x 2
Test niezależności chi2

Diagram D


Dyskretna zmienna zależna (ciągłe lub dyskretne zmienne niezależne)
Tak
Regresja logistyczna (logit)
Nie
Ciągła zmienna niezależna Ciągła zmienna zależna
Dwie lub więcej ciągłych zmiennych bez zmiennych niezależnych
Wstępnie ustalona zmienna niezależna
Nie
Rozkład normalny i jednorodność wariancji
Nie
Korelacja Spearmana
Tak
Liczba zmiennych ciągłych
2
Korelacja Pearsona
>2
Korelacja cząstkowa
Tak
Liczba zmiennych ciągłych niezależnych
Więcej niż jedna
Regresja wielokrotna
Jedna
Prosta analiza regresji

Rachunek prawdopodobieństwa

Prawdopodobieństwem (P) zwykło się nazywać szansę wystąpienia określonego zdarzenia (A). Ponieważ przyjęliśmy wyrażać taką szansę jako częstość pojawiania się A w zbiorowości wszystkich zdarzeń, jej miarą są liczby z przedziału od 0 do 1; 0 oznacza, że zdarzenie nie wystąpi nigdy, jedność – że wystąpi zawsze. Matematyczny aparat rachunku prawdopodobieństwa – stosowany na porządku dziennym, a często nawet bezwiednie, w naukach statystycznych – formalnie sprowadza się do podstawowych reguł logicznych.

Reguła addytywności (sumowania, *additive rule*) – kiedy mamy do czynienia z alternatywnym występowaniem cech lub zdarzeń – polega na sumowaniu prawdopodobieństw zdarzeń cząstkowych, zgodnie z ogólnym zapisem:

P (A lub B lub zarówno A jak i B) = P (A) + P (B) – P (zarówno A jak i B)

Reguła multiplikatywności (mnożenia, *multiplicative rule*) – kiedy badamy jednoczesne występowanie zdarzenia lub cechy – sprowadza się pod względem rachunkowym do iloczynu prawdopodobieństw zdarzeń cząstkowych, w zgodzie z zapisem:

P (A oraz B) = P (A) x P (B pod warunkiem, że wystąpiło A)

W przypadku zdarzeń warunkowych, kiedy wystąpienie zdarzenia B zależy od wystąpienia zdarzenia A (i *vice versa*), to P (A oraz B) = P (A) x P (B), jeżeli zdarzenia A i B są niezależne.

Rozkłady zmiennych*

W zależności od charakteru zmiennej (ciągła lub dyskretna) i/lub natury zjawiska przyrodniczego, które taka zmienna opisuje mamy do czynienia z różnymi typami rozkładu zmiennych. Każdy typ rozkładu ma swoje charakterystyczne cechy, które da się wyczytać z równania tzw. ***funkcji gęstości prawdopodobieństwa rozkładu*** (*probability density, density function*). Równanie to jest funkcją predykcyjną opisującą w jaki sposób „*zachowa się*" zmienna, to znaczy jakie jest prawdopodobieństwo, że będzie przyjmowała określone wartości. Czym wyższe jest takie prawdopodobieństwo, tym częściej wystąpi dana wartość zmiennej.

Ogólnie możemy sklasyfikować rozkłady na takie, które opisują zmienne ciągłe oraz takie, które wykorzystujemy do charakterystyki zmiennych dyskretnych. Do pierwszej grupy zaliczymy na przykład rozkład normalny, *t* Studenta, log-normalny, chi^2, Fishera-Snedecora, do drugiej: rozkład dwumianowy, Poissona, wykładniczy, Bernoulliego.

Poniżej podano krótkie charakterystyki wybranych typów rozkładów, z którymi mamy najczęściej do czynienia w naukach biomedycznych.

* Ten rozdział może być pominięty przez mniej zaawansowanych Czytelników bez szkody dla zrozumienia dalszych części opracowania.

Klasycznym rozkładem opisującym zmienną ciągłą jest opisany poniżej rozkład normalny.

Rozkład normalny

Rozkład normalny (o charakterystycznym kształcie krzywej dzwonowej, symetrycznej w stosunku do średniej) jest teoretycznym rozkładem prawdopodobieństwa powszechnie wykorzystywanym we wnioskowaniu statystycznym jako przybliżenie rozkładu z próby. Uważa się, że rozkład normalny jest dobrym modelem opisującym rzeczywisty rozkład zmiennej losowej, w sytuacji gdy:

1) istnieje silna tendencja do przyjmowania wartości położonych blisko środka rozkładu;
2) dodatnie i ujemne odchylenia od wartości środkowej występują z równym prawdopodobieństwem;
3) ze wzrostem wartości odchyleń ich liczebność gwałtownie spada.

Rozkład normalny posiada następującą funkcję gęstości:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x_i-\mu)^2/2\sigma^2} \quad \text{dla} \quad -\infty < x < \infty$$

gdzie:

μ jest średnią
σ jest odchyleniem standardowym
e jest stałą Eulera (2.718...)
π liczba pi (3.1415...).

Podstawowy mechanizm generowania rozkładu normalnego można wyobrazić sobie jako nieskończoną liczbę niezależnych zdarzeń losowych (dwumianowych), które generują wartości danej zmiennej. Na przykład, z dużym prawdopodobieństwem możemy stwierdzić, że występuje niemal nieograniczona liczba czynników genetycznych i środowiskowych, które wpływają na masę ciała u człowieka (wiele różnych genów, tryb życia, dieta, przebyte choroby, itp.). Toteż możemy spodziewać się, że w dużej liczebnie populacji masa ciała podlega rozkładowi normalnemu.

Rozkład t Studenta

To co różni ten rozkład od rozkładu normalnego, to kształt krzywej rozkładu zależny od liczebności (a w związku z tym i od liczby stopni swobody) próby. Prawdopodobieństwo, że zmienna przyjmie określoną wartość dla określonej liczebności próby określa funkcja gęstości rozkładu:

$$f(x) = \frac{\Gamma\left(\frac{v+1}{2}\right) * \left(\frac{1}{\sqrt{v*\pi}}\right) * \left(1+\frac{x^2}{v}\right)^{-\left(\frac{v+1}{2}\right)}}{\Gamma\left(\frac{v}{2}\right)}$$

gdzie:
v jest parametrem kształtu – liczbą stopni swobody
Γ jest funkcją gamma
π jest liczbą pi (3.1415...).

Rozkład chi²

Zmienna, która jest sumą kwadratów v niezależnych zmiennych losowych, z których każda podlega rozkładowi normalnemu posiada rozkład chi² (χ^2, *chi-squared distribution*) z liczbą stopni swobody równą v. W zastosowaniach statystycznych rozkład ten jest jednym z najczęściej stosowanych.

Posiada on następującą funkcję gęstości:

$$f(x) = \frac{1}{2^{v/2} * \Gamma\left(\frac{v}{2}\right)} * x^{\left(\frac{v}{2}-1\right)} * e^{-\frac{x}{2}}$$

dla $x > 0$, $n = 1, 2,...$
gdzie:
v jest liczbą stopni swobody
e jest stałą Eulera (2.718...)
Γ jest funkcją gamma (z argumentem α).

Rozkład Fishera-Snedecora

Wartość krytyczna $F = s_1^2 / s_2^2$ w metodach analizy wariancji, mówiąca ile razy wariancja wynikająca z różnic między średnimi przewyższa wariancję opisującą zmienność wewnątrz grup, podlega rozkładowi Fishera-Snedecora (*F (Snedecore's) distribution*).

Rozkład ten posiada następującą funkcję gęstości dla stopni swobody n licznika oraz ω mianownika:

$$f(x) = \frac{\Gamma\left(\frac{v+\varpi}{2}\right)}{\Gamma\left(\frac{v}{2}\right)\Gamma\left(\frac{\varpi}{2}\right)} * \left(\frac{v}{\varpi}\right)^{\frac{v}{2}} * x^{\left(\frac{v}{2}-1\right)} * \left[1 + \left(\frac{v}{\varpi} * x\right)\right]^{-\left(\frac{v+\varpi}{2}\right)}$$

dla $0 \le x < \infty$ oraz $v = 1, 2,...$; $\omega = 1, 2,...$
gdzie stopnie swobody:
v, ω są parametrami kształtu rozkładu, zaś
Γ jest funkcją gamma.

Jak można zauważyć z opisu, te cztery rozkłady (normalny, *t* Studenta, chi^2 i Fishera-Snedecora) mają wiele cech wspólnych, czego praktycznym następstwem jest to, że ich krytyczne wartości (a zatem i prawdopodobieństwa odpowiadające polom pod krzywą rozkładu) możemy swobodnie przeliczać. Krytyczne wartości testów: normalnego, *t* Studenta, Fishera-Snedecora i chi^2 są połączone zależnością:

$$z_{\alpha(2)} = t_{\alpha(2),\infty} = \sqrt{F_{\alpha(1),1,\infty}} = \sqrt{\chi^2_{\alpha,1}}$$

Wartości prawdopodobieństwa P można „*z grubsza*" oszacować wg równania:

$$P = \left[\frac{1 - \sqrt{1 - e^{-c^2}}}{2} \right]$$

gdzie $c = 0.806z(1 - 0.018z)$ dla z tak małych jak 0.2 oraz $c = z/(1.237 + 0.0249z)$ dla z tak małych jak 0.1.
Innym często stosowanym typem rozkładu jest opisany poniżej rozkład log-normalny.

Rozkład log-normalny

Rozkład tego typu jest często wykorzystywany do modelowania rozkładów zmiennych takich jak wysokość dochodów, wiek w momencie zawierania małżeństwa lub długość okresu przeżycia zwierząt w niewoli. Jeżeli x jest próbą pochodzącą z populacji o rozkładzie normalnym, to $y = e^x$ jest próbą o rozkładzie logarytmiczno-normalnym (*log-normal distribution*).

Rozkład logarytmiczno-normalny ma rozkład gęstości prawdopodobieństwa określony funkcją:

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} * e^{-\frac{(\log x - \mu)^2}{2\sigma^2}}$$

$0 < x < \infty$; $\mu > 0$; $\sigma > 0$
gdzie:
μ jest parametrem skali (odpowiada wartości średniej)
σ jest parametrem kształtu (odpowiada odchyleniu standardowemu)
e jest stałą Eulera (2.718...)
π jest liczbą pi (3.1415...).

Rozkład logarytmiczno-normalny ma zastosowanie w tych przypadkach, gdy rozkład zmiennej nie jest normalny (a jest na przykład prawoskośny), natomiast wartość logarytmu zmiennej losowej ma rozkład normalny. Tak jak rozkład normalny można sobie wyobrazić jako wynik dodawania dużej liczby niezależnych błędów, tak samo rozkład logarytmiczno-normalny można traktować jako wynik pomnożenia dużej liczby niezależnych błędów (czyli dodawania błędów logarytmicznych). Jeżeli zmienność cechy jakościowej jest pod wpływem dużej liczby losowych impulsów, które pojawiają się w czasie i wpływają proporcjonalnie do czasu, w którym się pojawiają, wówczas rozkład charakterystyki jakościowej będzie logarytmiczno-normalny. Na przykład rozmiar ciała, który jest wynikiem ciągłych procesów (etapów) wzrostowych, z których na każdy oddziałuje duża liczba niezależnych czynników, podlega rozkładowi logarytmiczno-normalnemu.

Wśród rozkładów, które wykorzystujemy często do charakterystyki zmiennych dyskretnych należy wymienić rozkłady: dwumianowy, Poissona, wykładniczy oraz Bernoulliego.

Rozkład wykładniczy

Jeżeli *t* oznacza czas pomiędzy kolejnymi zdarzeniami rzadkiego zjawiska, zachodzącego średnio *l* razy na jednostkę czasu, to *t* podlega rozkładowi wykładniczemu (*exponential distribution*) z parametrem λ. Przykłady zmiennych podlegających temu rozkładowi to np.: odstęp czasu pomiędzy przejazdami samochodów przez skrzyżowanie, czas bezawaryjnej pracy urządzeń elektronicznych lub czas między pojawianiem się ludzi w sklepie.

Rozkład wykładniczy posiada następującą funkcję gęstości:

$$f(x) = \lambda e^{-\lambda x}$$

dla $0 \leq x < \infty$; $\lambda > 0$
gdzie:
λ jest parametrem skali
e jest stałą Eulera (2.718...).

Rozkład dwumianowy

Rozkład tego typu jest wykorzystywany do opisu rozkładu zdarzeń dwumianowych, np. liczba kobiet i mężczyzn w próbie losowej pobranej w kilku grupach zawodowych lub liczba uszkodzonych elementów występujących w próbie 100 sztuk pobranych losowo.

Rozkład dwumianowy (*binomial distribution*) jest zdefiniowany jako:

$$f(x) = \frac{n!}{x!(n-x)!} * p^x * q^{n-x}$$

dla $x = 0, 1, 2, \ldots, n$
gdzie:
p oznacza prawdopodobieństwo sukcesu w każdej próbie
q oznacza prawdopodobieństwo porażki i jest równe $1-\pi$
n oznacza liczbę niezależnych prób.

Rozkład Poissona

Rozkład ten opisuje występowanie zdarzeń rzadkich, na przykład liczbę wypadków samochodowych na osobę, liczbę trafionych w rzucaniu kostką, liczbę usterek wytwarzanego produktu, ale też na przykład liczbę kropli deszczu, które spadły na jednostkę powierzchni, czy liczbę zbiorowisk wieloosobniczych nietoperzy zimujących w jaskini.

Opisuje go funkcja:

$$f(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

dla $x = 0, 1, 2, \ldots, 0 < \lambda$

gdzie:
λ jest wartością oczekiwaną x (średnią)
e oznacza stałą Eulera (2.718...).

Rozkład Bernoulliego (*Bernoulli binomial distribution)*

Rozkład tego typu najlepiej opisuje sytuacje, w których wynikiem doświadczenia może być sukces lub porażka, tak jak w przypadku rzutu monetą lub przy określaniu powodzenia lub niepowodzenia zabiegu operacyjnego.

Funkcja gęstości tego rozkładu jest zdefiniowana jako:

$$f(x) = p^x (1-p)^{1-x}$$

dla $x \in \{0.1\}$
gdzie:
p jest prawdopodobieństwem zajścia określonego zdarzenia.

W dalszej części tego opracowania dokładniej omówimy: 1) rozkład normalny, mający odniesienie do zmiennych ciągłych, i na którego charakterystyce oparta jest większość testów parametrycznych, 2) rozkład dwumianowy, opisujący rozkłady proporcji zmiennych dyskretnych, oraz 3) rozkład Poissona, opisujący częstość występowania zjawisk rzadkich.

Rozkład normalny

Rozkład normalny, który charakteryzuje się dzwonowatym kształtem krzywej symetrycznej w stosunku do wartości centralnej (średniej), jest jednym z najpowszechniej wykorzystywanych teoretycznych rozkładów prawdopodobieństwa we wnioskowaniu statystycznym, jako przybliżenie rozkładu z próby. Rozkład ten jest dobrym przybliżeniem do rzeczywistości, i stanowi dobry model dla rozkładu zmiennej losowej, wszędzie tam, gdzie:

- zmienna ma silną tendencję do przyjmowania wartości położonych blisko środka rozkładu,
- odchylenia dodatnie i ujemne od wartości centralnej są z dużym prawdopodobieństwem równe,
- duże odchylenia od wartości centralnej są bardzo nieliczne, a liczebność odchyleń spada gwałtownie w miarę wzrostu ich wartości.

Rozkłady zmiennych występujących w świecie rzeczywistym, na których wartości wpływa niemal nieograniczona liczba czynników determinujących (genetycznych i środowiskowych), aproksymują silnie do rozkładu normalnego. Do takich zmiennych należą na przykład wzrost, masa ciała, temperatura, ciśnienie krwi, stężenie hemoglobiny. Zmienną, która nie spełnia rozkładu normalnego jest na przykład dochód. Niekiedy transformacja wartości zmiennych sprzyja normalizacji rozkładu, na przykład transformacja logarytmiczna normalizuje z reguły rozkład prawoskośny. Więcej informacji na temat transformacji danych znajdzie Czytelnik w Rozdziale *„Transformacja danych – sposoby „normalizacji" rozkładu"* (str. 97).

O przystawalności rozkładu danych doświadczalnych do rozkładu normalnego wnioskujemy najczęściej przeprowadzając jeden z testów normalności (np. test zgodności chi^2, test Kolmogorowa-Smirnowa czy test W Shapiro-Wilka), które pozwalają obliczyć prawdopodo-

bieństwo tego, że próba pochodzi z populacji o rozkładzie normalnym. Procedury te omówiono w *„Części II – Uzupełnienia, przykłady i zadania"*.

Rozkład normalny jest taki ważny nie tylko dlatego, że jest on dobrym doświadczalnym odzwierciedleniem większości zmiennych spotykanych w przyrodzie, ale także dlatego, że założenie normalności jest wymaganiem większości stosowanych w naukach przyrodniczych testów i metod analizy statystycznej. Jest tak dlatego, że większość z tych testów albo bezpośrednio wywodzi się z rozkładu normalnego, albo jest z nim związana (np. test *t* Studenta, test Fishera-Snedecora czy test chi^2). Najczęściej spełnianie normalności rozkładu zmiennych jest warunkiem stosowania takich testów. Problem powstaje, gdy usiłujemy zastosować test oparty na założeniu o normalności do zmiennych, które nie posiadają rozkładu normalnego. W takim przypadku możemy albo zastosować testy niewymagające założenia o normalności (tak zwane testy niezależne od rozkładu lub testy nieparametryczne) lub posłużyć się testami opartymi o normalność pod warunkiem, że dysponujemy dostatecznie liczną próbą. Pierwsze rozwiązanie jest z reguły niedogodne ze względu na małą moc takich testów i ich nieelastyczność w formułowaniu wniosków. Druga możliwość opiera się na zasadzie – zwanej ***centralnym twierdzeniem granicznym*** (*central theorem*) – dzięki której testy oparte na rozkładzie normalnym posiadają tak wielkie znaczenie. Mówi ona, że:

- średnia wszystkich możliwych średnich prób losowanych z populacji ogólnej jest równa średniej populacji ogólnej;

$$\overline{X_{\overline{X}}} = \mu$$

- odchylenie standardowe wszystkich możliwych średnich prób losowanych z populacji ogólnej jest równe odchyleniu standardowemu populacji ogólnej podzielonemu przez pierwiastek z liczebności próby;

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

- w miarę jak wzrasta liczność próbki, rozkład statystyki testowej z próby zbliża się do rozkładu normalnego, nawet jeśli zmienna, którą mierzymy, nie posiada rozkładu normalnego.

Ponieważ stosowanie testów opartych na charakterystyce rozkładu normalnego jest takie wygodne, to często ulegamy pokusie, aby wykorzystywać takie testy bez uprzedniego sprawdzania warunku normalności, albo nawet wtedy, gdy warunek taki jest naruszony. Powstaje wtedy wątpliwość, czy konsekwencje niespełnienia założenia o normalności są poważne i skąd wiemy o naruszeniu tego założenia. Empiryczna ocena typów i wielkości błędów popełnianych w przypadku niespełnienia założeń niezbędnych do zastosowania specyficznych testów (dokonywana często przy użyciu tzw. metody Monte Carlo) wskazuje, że konsekwencje złamania założenia o normalności nie są na ogół takie poważne, jak sądzono wcześniej. Może to zachęcać do szerszego stosowania metod statystycznych zależnych od rozkładu normalnego, ale nie do całkowitego zaniechania troski o spełnienie założenia o normalności rozkładów badanych zmiennych.

Rozkład normalny posiada kilka niezwykle użytecznych cech charakterystycznych. Przede wszystkim z kształtu i symetryczności rozkładu wynika, że około 68% wszystkich obserwacji trafia do przedziału $\mu \pm 1\sigma$ (w praktyce oznacza to $\bar{x} \pm s$), a przedział średniej ± 2 odchylenia standardowe obejmuje prawie 95% przypadków. Czyli w rozkładzie normal-

nym wartości mniejsze od średniej niż –2SD i większe niż +2SD zdarzać się mogą z częstością równą lub mniejszą niż 5%. Odpowiednio, wartości z przedziału $\mu \pm 3\sigma$ spotykamy z prawdopodobieństwem około 99.5%, zaś te z przedziału $\mu \pm 4\sigma$ wystąpią w ponad 99.9% przypadków. Stąd wyprowadza się praktyczną interpretację odchylenia standardowego (zobacz *„Miary rozproszenia"*).

Szczególnym przypadkiem rozkładu normalnego jest ***rozkład normalny standaryzowany*** (SND), gdy średnia wynosi 0, zaś wartość odchylenia równa się 1.

$$SND, z = \frac{x - \mu}{\sigma}$$

Sens statystyczny wartości z, która jest niczym innym jak wielokrotnością odchylenia standardowego, o jaką obserwowana wartość zmiennej różni się od średniej dla populacji generalnej, omówiono w dalszej części opracowania (zobacz też omówienie przykładu 16 w *„Części II – Uzupełnienia, przykłady i zadania"*).

Kształt rozkładu i normalność rozkładu

Dla rozkładu normalnego nie ma znaczenia w jakich jednostkach jest wyrażona zmienna. Zmiana wartości średniej powoduje przesunięcie krzywej rozkładu w lewo lub w prawo, podczas gdy zmiana odchylenia standardowego zmienia wysokość lub szerokość krzywej Gaussa, czyli wpływa na kształt rozkładu. Kształt rozkładu jest charakterystyczną cechą zmiennej: informuje on o częstości występowania różnych wartości tej zmiennej w różnych obszarach jej zmienności.

$$f_i = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x_i - \mu)^2 / 2\sigma^2}$$

gdzie:
f_i oznacza częstość obserwacji zmiennej x_i, zaś
μ oraz σ średnią i odchylenie standardowe populacji.

Parametrami kształtu rozkładu są skośność i kurtoza.

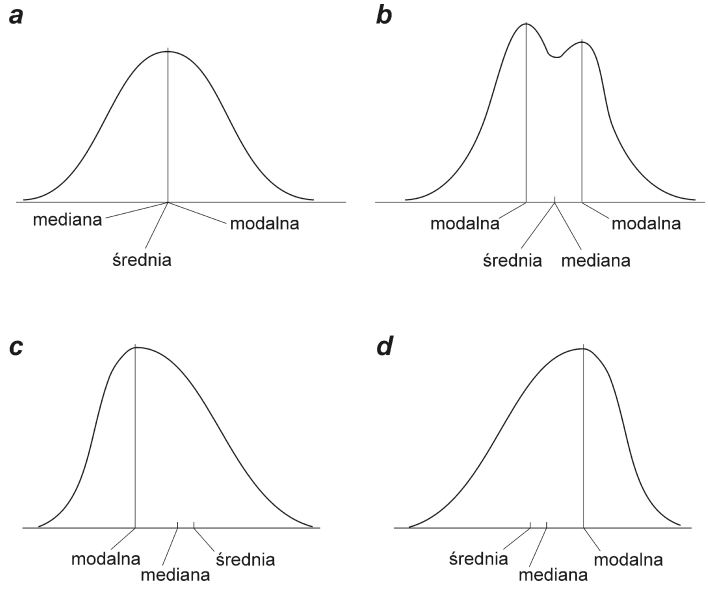
Skośność (*skeweness*) charakteryzuje odchylenie rozkładu od symetrii. Jeśli wartość skośności γ ($\kappa_3 = \frac{(x_i - \mu)^3}{N}$) po standaryzacji ($\gamma_1 = \frac{\kappa_3}{\sigma^3}$) jest wyraźnie różna od zera, wówczas dany rozkład jest asymetryczny; taki rozkład odstaje od charakterystyki rozkładu normalnego, gdyż rozkład normalny jest doskonale symetryczny. Dla rozkładu lewoskośnego (*left-skewed distribution*) $\gamma_1 < 0$, dla prawoskośnego (*right-skewed distribution*) $\gamma_1 > 0$ (Ryc. 1).

Dla rozkładu normalnego wartości średniej, mediany i modalnej są identyczne. W przypadku rozkładów lewoskośnych średnia jest mniejsza od mediany i obie te wartości leżą na lewo od modalnej. W rozkładach prawoskośnych największą wartość przyjmuje średnia, zaś najniższą modalna.

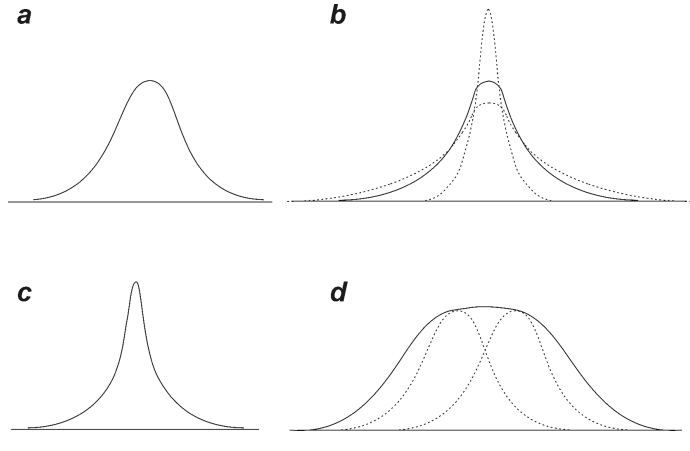
Kurtoza (*curtosis*), wyliczana ze wzoru:

$$\kappa_4 = \frac{(x_i - \mu)^4}{N}, \text{ po standaryzacji } \gamma_2 = \frac{\kappa_4}{\sigma^4} - 3,$$

mierzy spiczastość rozkładu. Jeżeli kurtoza (miara spłaszczenia rozkładu) jest wyraźnie różna od zera, wówczas rozkład jest albo bardziej spłaszczony niż rozkład normalny albo bardziej wysmukły (Ryc. 2). Kurtoza rozkładu normalnego wynosi dokładnie 0, wartości $\gamma_2 > 0$ charakteryzują rozkłady leptokurtyczne (wysmukłe), a $\gamma_2 < 0$ rozkłady platykurtyczne



Ryc. 1. Typy rozkładów: a) normalny, b) dwumodalny, c) prawoskośny, d) lewoskośny.



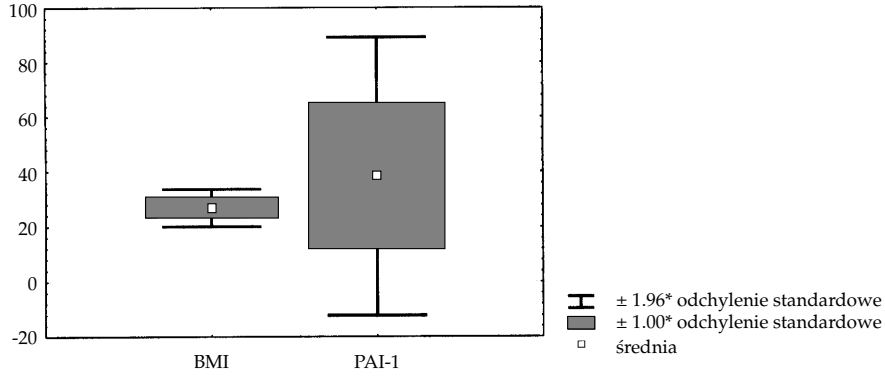
Ryc. 2. Kurtoza rozkładu: a) rozkład mezokurtyczny (normalny), b) złożony rozkład dwóch populacji, dla których $\bar{x}_1 \cong \bar{x}_2$ oraz $s_1 >> s_2$, c) rozkład leptokurtyczny (wysmukły), d) rozkład platykurtyczny (spłaszczony).

(spłaszczone). Rozkład platykurtyczny jest szczególnym przypadkiem rozkładu dwumodalnego (*bimodal distribution*) (Ryc. 1), to znaczy takiego, który posiada dwa maksima (Ryc. 2d). Taki objaw wskazuje, że próba nie jest jednorodna i jej obserwacje pochodzą z dwóch różnych populacji, z których każda posiada rozkład normalny. Taka ewidentna cecha powinna nas skłonić do rozseparowania próby na dwie subpopulacje i osobną charakterystykę każdej z nich. Typ rozkładu leptokurtycznego może wskazywać na występowanie dwóch niezależnych subpopulacji normalnych o zbliżonych wartościach średniej i różnych wariancjach (Ryc. 2b).

Przedział ufności dla średniej i wnioskowanie o wartości średniej dla populacji ogólnej

Jesteśmy najczęściej zainteresowani wartościami różnych statystyk próby badanej (takich jak na przykład średnia czy odchylenie standardowe) tylko o tyle, o ile informują nas one o stanie faktycznym w populacji generalnej. Zauważyliśmy poprzednio, że najczęściej nie znamy ani wartości średniej ani zmienności całej populacji ogólnej. Możemy jedynie z określonym prawdopodobieństwem i przybliżeniem szacować te wartości na podstawie określania parametrów próby losowej. W jaki sposób możemy wykorzystać estymowaną dla próby losowej wartość średniej i błędu standardowego w celu określenia najbardziej prawdopodobnej wartości średniej rzeczywistej? Okazuje się, że wartość średnia jest szczególnie użyteczną miarą tendencji centralnej rozpatrywanej zmiennej, jeśli jest podawana wraz ze swym przedziałem ufności (*confidence interval*). Częstą formą graficznej prezentacji miary centralnej, miary rozrzutu oraz przedziału ufności na jednym wykresie jest wykres ramkowy (tzw. wykres *„pudełka z wąsami"*, *box and whisker plot*) (Ryc. 3).

Z poprzedniego rozdziału wiemy, że przy spełnionym warunku normalności rozkładu 95% wszystkich średnich cząstkowych populacji losowych będzie leżało w obszarze objętym zakresem wyznaczonym przez wartość dwóch odchyleń standardowych na lewo i na prawo od średniej rzeczywistej. Średnia rzeczywista populacji, m będzie równa uśrednionej



Ryc. 3. Przykładowy wykres ramkowy ukazujący średnią ± SD oraz 95% przedział ufności dla dwóch zmiennych: wskaźnika masy ciała (BMI) oraz stężenia inhibitora tkankowego aktywatora plazminogenu (PAI-1) w osoczu w grupie pacjentów z chorobą wieńcową.

wartości wszystkich średnich losowych, a odchylenie standardowe będzie równe błędowi standardowemu średnich losowych. Estymacja taka jest usprawiedliwiona pod warunkiem dużej liczebności próby (powyżej 60), gdyż wtedy charakterystyka rozkładu próby jest bardzo bliska charakterystyce rozkładu normalnego nawet wtedy, gdy próba nie posiada *de facto* rozkładu normalnego. Dla takich prób odchylenie standardowe, s, jest dobrym odzwierciedleniem odchylenia populacji, σ. W przypadku rozkładu normalnego możemy stwierdzić, że 95% wszystkich średnich prób losowych leży w zakresie 1.96 odchyleń od średniej populacji, μ, czyli prawdopodobieństwo tego, że określona średnia losowa leży w obszarze objętym przez 1.96 odchyleń z każdej strony od wartości średniej dla populacji wynosi 95%. W praktyce właściwość ta jest stosowana, aby na podstawie średniej próby oraz błędu standardowego oszacować zakres, wewnątrz którego leży średnia rzeczywista. Dla wartości prawdopodobieństwa 95% zakres ten wynosi od $\bar{x} - 1.96 * SE$ do $\bar{x} + 1.96 * SE$ i nazywany jest 95% przedziałem ufności dla średniej (95% *CI*), zaś wartości od $\bar{x} - 1.96 * SE$ do $\bar{x} + 1.96 * SE$ określane są jako granice przedziału ufności (*confidence interval, CI*). Możemy zapisać, że:

$$95\%\ CI\ (\textit{duże próby}) = \bar{x} \pm (1.96 \times s / \sqrt{n}) .$$

Ponieważ różnym wartościom prawdopodobieństwa odpowiadają różne wartości dystrybuanty* rozkładu normalnego, postać ogólna powyższego równania wygląda następująco:

$$CI\ (\textit{duże próby}) = \bar{x} \pm (z' \times s / \sqrt{n}) .$$

W przypadku mało licznych prób poprawność powyższych rozważań może być podważona w dwóch przypadkach. Po pierwsze, odchylenie standardowe próby, s, może nie odzwierciedlać właściwie odchylenia populacji ogólnej, σ (zobacz Rozdział „*Błąd standardowy i precyzja określania wartości średniej próby*"). Po drugie, jeżeli rozkład populacji nie jest normalny, to rozkład średnich losowych może także nie być normalny. Drugi przypadek jest szczególnie istotny, gdy liczebność jest mniejsza niż kilkanaście obserwacji, a rozkład odstaje znacznie od rozkładu normalnego (zobacz *centralne twierdzenie graniczne*). Czym większa liczebność próby, tym większe prawdopodobieństwo bliskiej aproksymacji rozkładu próby do rozkładu granicznego. Graniczna wielkość próby zależy od tego, jak dalece dany rozkład odstaje od charakterystyki rozkładu normalnego, ale w praktyce dla większości sytuacji przyjmuje się liczebność większą od 15 przypadków.

Z powyższych względów, w praktyce borykamy się najczęściej jedynie z pierwszym problemem, kiedy zmienność próby nie jest właściwą reprezentacją zmienności populacji, czyli kiedy s różni się znacząco od σ. W rzeczywistości bardzo rzadko znamy faktyczną wartość σ, i jeżeli nie możemy z różnych względów zastąpić jej wartością s próby badanej, nie powinniśmy stosować rozkładu normalnego do szacowania przedziałów ufności. W przypadkach takich wykorzystujemy tzw. ***rozkład t Studenta***, który może być stosowany zawsze z wyjątkiem przypadków bardzo silnie zaznaczonego niespełnienia normalności rozkładu.

---

* W sensie matematycznym wartość dystrybuanty (*cumulative frequency, distribuant*) w punkcie x definiujemy jako prawdopodobieństwo tego, że zmienna losowa x_i przyjmie wartości mniejsze lub równe wartości x (zobacz przykłady w „*Części II – Uzupełnienia, przykłady i zadania*").

Rozkład *t* Studenta ma charakterystykę bardzo zbliżoną do rozkładu normalnego: ma dzwonowaty symetryczny kształt, wartość średnia wynosi 0 – podobnie jak w standaryzowanym rozkładzie normalnym – ale jest on nieco bardziej skośny (rozciągnięty na boki) od rozkładu normalnego, szczególnie przy niskich liczebnościach próby. Kształt rozkładu *t* zależy istotnie od liczby stopni swobody (*d.f.* = *n*–1) (a więc i od liczebności): czym mniejsza liczba stopni swobody, tym większe rozciągnięcie na boki. Punkty krytyczne (*critical value, percentage point*) tego rozkładu odczytujemy zawsze dla wypadkowych wartości prawdopodobieństwa (pola pod krzywą) i liczby stopni swobody.

Przedział ufności dla małych prób szacowany jest w oparciu o wartości krytyczne testu *t* dla liczby stopni swobody (*n*–1):

$$CI\ (\textit{małe próby}) = \bar{x} \pm (t' \times s / \sqrt{n})$$

Dla małej liczby stopni swobody wartości krytyczne rozkładu *t* (*t'*) są zdecydowanie wyższe niż odpowiednie wartości *z'* rozkładu normalnego, ponieważ w sytuacjach takich wysokie jest prawdopodobieństwo, że odchylenie standardowe próby, *s*, jest słabą reprezentacją odchylenia standardowego populacji, σ, i niepewność takiej estymacji sprawia, że przedział ufności jest znacząco szerszy niż przy estymacji dla rozkładu normalnego przy znanej wartości σ. W miarę wzrostu liczebności próby, a więc i liczby stopni swobody, wartość *s* staje się coraz bardziej reprezentatywna dla populacji i coraz bliższa σ, a rozkład *t* staje się prawie identyczny jak rozkład normalny.

W przypadkach gdy normalność rozkładu nie jest spełniona i odstępstwo od normalności jest bardzo silnie zaznaczone, można wykorzystać różne procedury transformacji danych, które służą normalizacji rozkładu, zastosować procedury (testy) odrzucania wyników niepewnych, lub też stosować metody nieparametryczne. Pierwsze czy drugie podejście jest zdecydowanie lepsze niż ostatnie, gdyż nawet pomijając fakt, że metody liczenia nieparametrycznych przedziałów ufności są niezwykle skomplikowane, to należy zawsze pamiętać o korzyściach, walorach, a także wygodzie stosowania testów opartych na normalności rozkładu.

W Tabeli 1 zebrano metody obliczania przedziałów ufności dla różnych liczebności i typów rozkładu próby badanej.

Tab. 1. Zalecane metody obliczania przedziału ufności.

	rozkład próby	
liczebność próby	**w przybliżeniu normalny**	**silnie odbiegający od normalnego***
odchylenie standardowe populacji σ nieznane		
60 lub więcej	$\bar{x} \pm (z' \times s / \sqrt{n})$	$\bar{x} \pm (z' \times s / \sqrt{n})$
mniej niż 60	$\bar{x} \pm (t' \times s / \sqrt{n})$	testy nieparametryczne
odchylenie standardowe populacji σ znane		
15 lub więcej	$\bar{x} \pm (z' \times \sigma / \sqrt{n})$	$\bar{x} \pm (z' \times \sigma / \sqrt{n})$
mniej niż 15	$\bar{x} \pm (z' \times \sigma / \sqrt{n})$	testy nieparametryczne

* pożądane jest poddać dane transformacji tak, aby ich rozkład odpowiadał bardziej charakterystyce rozkładu normalnego.

Podsumowując, przedział ufności dla wartości średniej określa zakres wartości wokół średniej, co do którego spodziewamy się, że prawdziwa (tzn. ta charakterystyczna dla populacji) wartość średnia mieści się w nim z określonym prawdopodobieństwem. Jeśli na przykład w naszej próbie średnia wynosi 20, a dolna i górna granica przedziału ufności wynoszą z prawdopodobieństwem 95% (czyli na poziomie istotności $\alpha = 0.05$) odpowiednio 15 i 25, to możemy wnioskować, że prawdopodobieństwo tego, iż średnia wartość w populacji jest zawarta w przedziale od 15 do 25 wynosi 95%. Gdybyśmy zwiększyli to prawdopodobieństwo (czyli zmniejszyli istotność), wówczas przedział uległby poszerzeniu zwiększając tym samym pewność oceny (i na odwrót). Jak uczy doświadczenie, im mniej konkretna jest prognoza (tzn. im szerszy przedział ufności), tym bardziej możemy być pewni, że się ona sprawdzi. Należy także pamiętać, że wielkość przedziału ufności zależy od wielkości próby oraz od zmienności cechy badanej. Im większa jest próba, tym bardziej wiarygodna jest ocena wartości średniej, natomiast im większa zmienność cechy, tym ocena średniej jest mniej wiarygodna. Obliczanie przedziałów ufności opiera się na założeniu, że rozkład zmiennej w populacji generalnej jest normalny. Ocena może nie być dokładna, jeśli to założenie nie jest spełnione, chyba, że próba jest duża.

Kierując się wartościami dystrybuanty dla różnych punktów krytycznych rozkładu normalnego, przyjęto umownie, aby obserwacje o wartościach odstających od średniej o więcej niż dwa odchylenia standardowe traktować jako ostrzegawcze, zaś te które różnią się o więcej niż 3 odchylenia od średniej – jako wątpliwe.

Rozkład dwumianowy

Jeżeli dla każdego osobnika w próbie zmienna przyjmuje jedną z dwóch możliwych wartości (A lub B, albo 0 lub 1, np. dodatni lub ujemny wynik testu, zgon lub przeżycie), to zmienna taka nazywa się zmienną binarną, zaś jej rozkład – rozkładem dwumianowym. Liczbę osobników, dla których wartością zmiennej jest jeden z takich alternatywnych wyborów (r), do całkowitej liczebności próby (n) nazywamy proporcją (p) cechy A w próbie:

$$p = \frac{r}{n}$$

Zapis ten odczytujemy: prawdopodobieństwo (p), że cecha A pojawi się r razy wśród n osobników. Proporcja cechy A charakteryzuje się oczywiście jakąś zmiennością w populacji i różne próby losowe będą posiadały różniące się wartości proporcji p. Wartości te podlegają rozkładowi dwumianowemu i mogą być obliczone na podstawie ogólnej liczebności próby n oraz proporcji cechy A w populacji (π), która wyraża prawdopodobieństwo, że dowolny wylosowany przypadkowo z tej populacji osobnik posiada cechę A.

Prawdopodobieństwo, że A wystąpi dokładnie r razy w próbie o liczebności n, gdy prawdopodobieństwo wystąpienia A u dowolnego osobnika wynosi π, obliczymy według równania:

$$P(r\ A) = \frac{n!}{r!(n-r)!}\pi^r(1-\pi)^{n-r}$$

Pierwsza część tego równania reprezentuje liczbę wszystkich możliwych kombinacji, w jakich cecha A może być spotkana r razy w populacji o liczebności n, zaś druga część wyraża prawdopodobieństwo każdej z tych możliwych kombinacji. Tą pierwszą część równania możemy rozpisać jako:

$$\frac{n!}{r!(n-r)!} = \frac{n\times(n-1)\times(n-2)\times....\times(n-r+1)}{r\times(r-1)\times....\times3\times2\times1}$$

Na Rycinie 4 przedstawiono przykłady rozkładu dwumianowego dla różnych liczebności próby n oraz różnych wartości prawdopodobieństwa π wystąpienia w populacji cechy A. Rozkłady te są zilustrowane dla różnych obserwowanych wartości r.

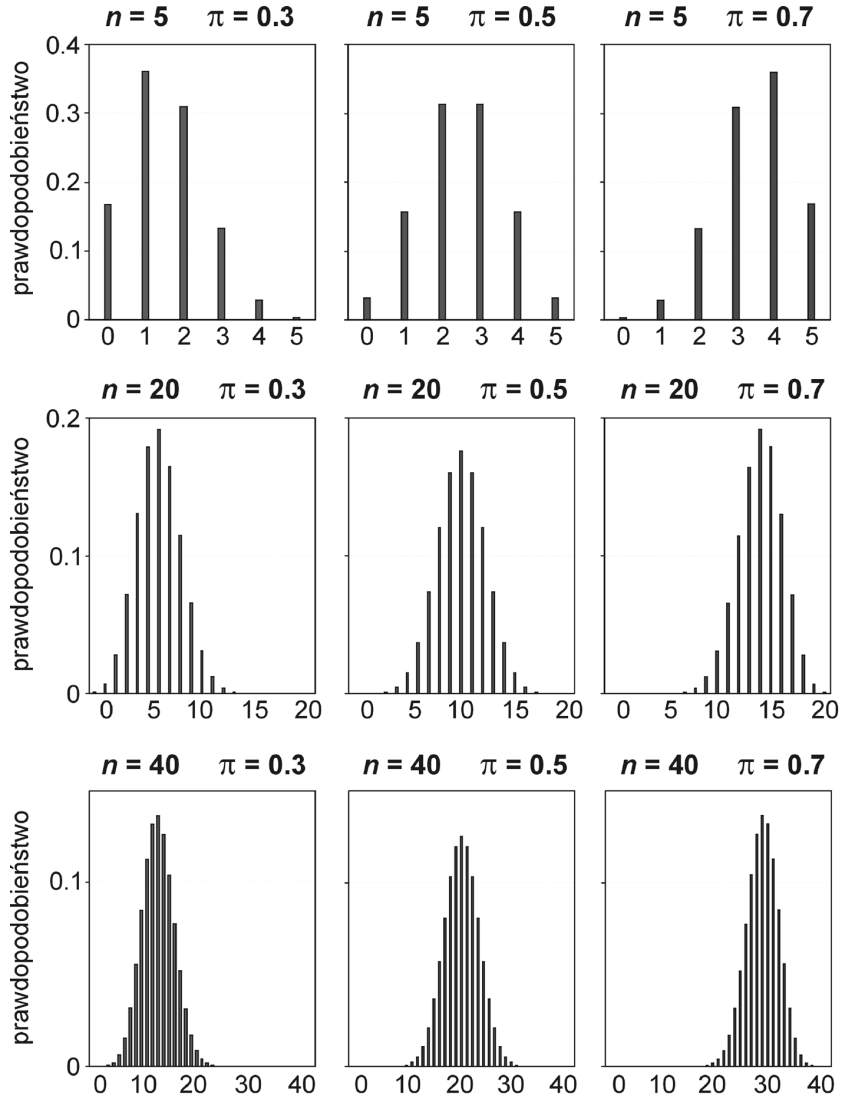
Zamiast obserwowanych wartości r możemy zastosować wartość proporcji p dla odpowiednich r. Dla próby o liczebności n, liczba wystąpień cechy r będzie przyjmować wartości 1, 2, 3, 4, 5.... n, zaś odpowiednie proporcje będą wynosić: $1/n, 2/n, 3/n, 4/n, 5/n$ ...1. Zauważmy, że chociaż p jest ułamkiem, rozkład możliwych wartości p jest dyskretny a nie ciągły, ponieważ w określonej próbie może wystąpić jedynie ograniczona liczba możliwych wariantów p.

Z uwagi na ten dyskretny rozkład wartości obserwowanych r cechy A (lub proporcji), wartość średnia rozkładu dwumianowego cechy A jest równa średniej tej cechy w populacji, natomiast jej zmienność jest wyrażona przez wartość błędu standardowego, który określa jak dobrze wartość obserwowana próby odzwierciedla wartość populacji ogólnej (Tab. 2).

Tab. 2. Średnia i błąd standardowy populacji dla liczby wartości obserwowanych, proporcji oraz częstości cechy A (wyrażonej w procentach).

	wartość obserwowana	**średnia populacji**	**błąd standardowy**
liczba wystąpień cechy A	r	$n\pi$	$\sqrt{\{n\pi(1-\pi)\}}$
proporcja cechy A	$p = r/n$	π	$\sqrt{\{\pi(1-\pi)/n\}}$
procent cechy A	$100p$	100π	$100\sqrt{\{\pi(1-\pi)/n\}}$

Uwaga: Na charakterystyce rozkładu dwumianowego oparte są testy istotności dla proporcji omówione w dalszej części tego opracowania.



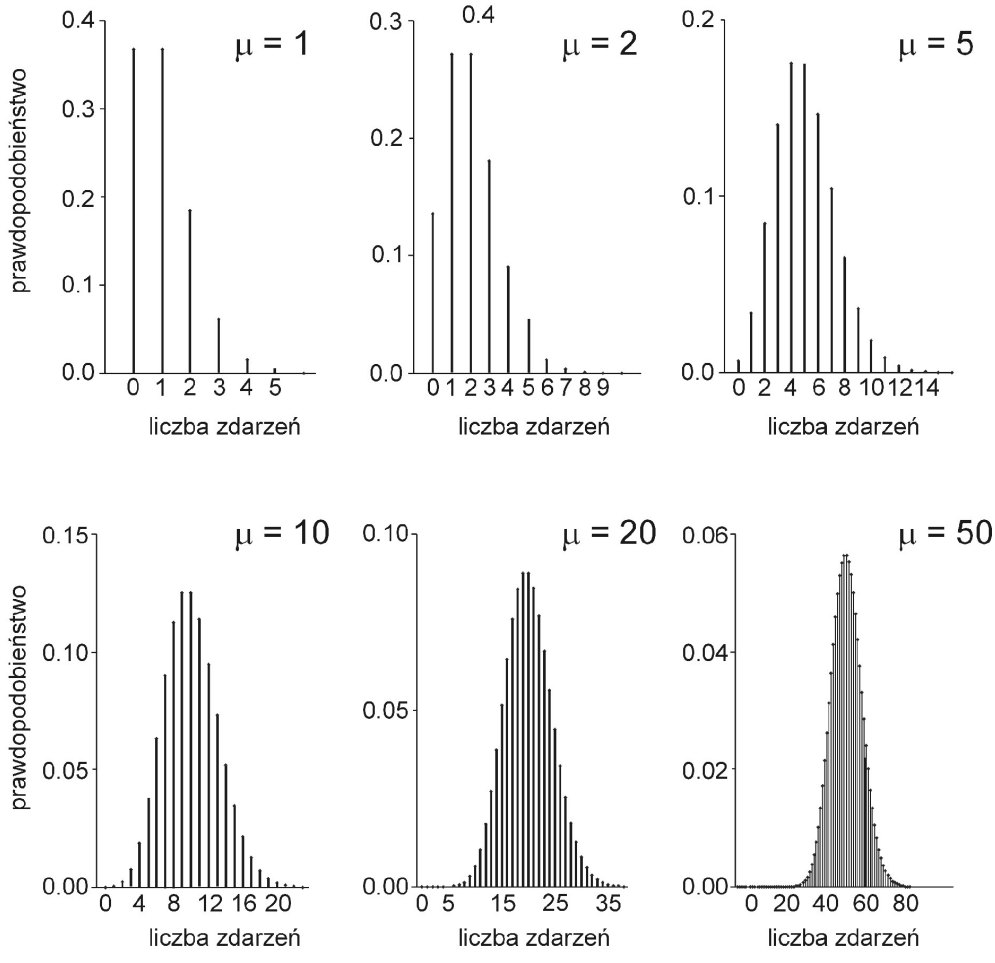
Ryc. 4. Rozkłady dwumianowe dla różnych liczebności próby n oraz różnych wartości prawdopodobieństwa π. Wartości obserwowane r zaznaczono na osi odciętych.

Rozkład Poissona

Rozkład ten opisuje występowanie zdarzeń, zwykle pojawiających się niezbyt często, w określonym interwale czasu. Warunkiem, aby ten typ rozkładu dobrze odzwierciedlał rzeczywistość jest, aby zdarzenia takie były niezależne od siebie i pojawiały się w sposób czysto przypadkowy. Wyobraźmy sobie pole w kształcie kwadratu, które podzielone jest na 64 mniejsze pólka – tak jak na szachownicy. Na pole takie spadają krople deszczu, przy czym oczywiście łatwo sobie wyobrazić, że w ciągu mierzalnego przedziału czasu na każde z 64 mniejszych spadnie nieidentyczna liczba kropli deszczu: na niektóre pólka spadnie po jednej kropli, na niektóre po dwie, rzadko po trzy, jeszcze rzadziej po 4 krople, a zupełnie sporadycznie wystąpi sytuacja, kiedy na pólko spadnie na przykład więcej niż 10 kropli. Przyjdzie nam to łatwiej, jeżeli wyobrazimy sobie, że jest to nieintensywny, przelotny deszcz – nie zaś rzęsista ulewa – oraz że naszej obserwacji dokonujemy w niedługim czasie. Rozkład prawdopodobieństw, że na każde z pojedynczych pól spadnie wzrastająca liczba kropli deszczu w danym przedziale czasu najlepiej opisuje właśnie rozkład Poissona. Zauważmy, że opisane powyżej zjawisko dobrze obrazuje ograniczenia tego rozkładu. Powinny to być zdarzenia rzadkie, pojawiające się sporadycznie, gdyż w miarę wzrostu częstości zdarzenia i wydłużania czasu obserwacji będziemy się przybliżali do sytuacji, kiedy liczba elementarnych zdarzeń (w naszym przypadku liczba kropli deszczu, które spadły na każde z 64 pól) będzie coraz bardziej wyrównana.

Inne przykłady zjawisk przyrodniczych, czy tych przykuwających naszą uwagę w naukach przyrodniczych i biomedycznych, gdzie rozkład Poissona znajduje zastosowanie, to liczba rozpadów pierwiastka promieniotwórczego wykryta za pomocą licznika scyntylacyjnego w krótkim, na przykład 2-3 minutowym, przedziale czasu, liczba okazów stonki ziemniaczanej znaleziona na liściach sadzonek ziemniaka, liczba obiektów znalezionych w jednostkowej przestrzeni (pod warunkiem, że obiekty te są rozmieszczone w sposób przypadkowy i niezależny jeden od drugiego), np. liczba komórek pasożyta przypadająca na zainfekowaną komórkę, czy liczba przypadków zachorowań (infekcji) w badanej grupie osób. Ten ostatni przykład nawiązuje do metod analizy częstości wystąpienia choroby (zachorowalności) omówionych w dalszej części tego opracowania (zobacz *„Metody wykorzystywane w badaniach populacyjnych"*). Zwróćmy uwagę na istotność podstawowego wymagania tego rozkładu, to znaczy przypadkowego i niezależnego rozmieszczenia zdarzeń/obiektów. Na przykład, liczba jaj pasożyta *Schistosoma mansoni* w preparatach próbek kału nie będzie podlegała temu rozkładowi, gdyż jaja te mają tendencję do zlepiania się i skupiania, czyli ich rozkład nie będzie niezależny. Podobnie, powinniśmy uwzględnić efekt skupiania w czasie i przestrzeni w przypadku niektórych chorób (np. rozkład zachorowalności na białaczkę u dzieci w regionie elektrowni jądrowej), gdyż może to zaburzyć charakterystykę rozkładu Poissona. Aby upewnić się czy taki typ rozkładu nie jest naruszony, korzystamy z procedur sprawdzania typu rozkładu, opisanych w *„Części II – Uzupełnienia, przykłady i zadania"*.

Pod względem rachunkowym funkcja gęstości rozkładu Poissona opisuje prawdopodobieństwo wystąpienia określonej liczby x zdarzeń w przedziale czasu, przestrzeni (lub regionie) i zależy tylko od jednego parametru μ, który jest średnią liczbą zdarzeń w przedziałach czasu o tej samej długości (lub w jednakowych regionach albo obszarach przestrzeni):



Ryc. 5. Rozkłady Poissona dla różnych wartości μ.

$$P(x \text{ zdarzeń}) = \frac{e^{-\mu}\mu^{x}}{x!}.$$

Ponieważ zarówno 0! jak i μ^0 są z definicji równe jedności, to prawdopodobieństwo, że zdarzenie pojawi się 0 razy (czyli nie wystąpi wcale) wynosi $e^{-\mu}$.

Błąd standardowy dla wielu niezależnych zdarzeń równy jest pierwiastkowi średniej:

$$SE = \sqrt{\mu}$$

i szacowany jest w praktyce jako $\sqrt{x}$.

Wartość parametru μ decyduje o kształcie rozkładu Poissona (Ryc. 5). Kształt rozkładu Poissona jest bardzo prawoskośny dla niskich wartości średnich μ, kiedy należy się liczyć z wysokim prawdopodobieństwem niewystąpienia zdarzenia. Dla wzrastających μ (μ >10) rozkład Poissona aproksymuje do rozkładu normalnego.

Wykorzystanie rozkładu Poissona w analizie częstości

Rozkład Poissona może być z powodzeniem wykorzystany do analizy częstości wystąpienia choroby (zachorowalności) (zobacz *„Metody wykorzystywane w badaniach populacyjnych"*).

W analizie występowania (rzadkich) zdarzeń w czasie możemy wykorzystywać charakterystykę rozkładu Poissona, jeśli tylko rejestrowane zdarzenia pojawiają się niezależnie od siebie i są w sposób przypadkowy rozłożone w czasie. Pamiętamy, że średnią liczbę zdarzeń w kolejnych obserwacjach wyrażamy jako μ, zatem możemy zapisać, że średnia liczba zdarzeń μ pojawiająca się w jednostce czasu t będzie wynosić:

$$\lambda = \frac{\mu}{t} \quad \text{oraz} \quad SE = \frac{\sqrt{\mu}}{t} = \sqrt{\frac{\lambda}{t}}$$

Jak widać, wartość błędu standardowego maleje w miarę wydłużania czasu obserwacji, jako że wartość λ będzie w przybliżeniu taka sama dla krótkich i długich okresów obserwacji.

Rozdział 4

Testy istotności dla pojedynczej próby lub dwóch prób

Ta grupa metod obejmuje testy umożliwiające odpowiedź na pytanie, czy średnia próby badanej jest zgodna z hipotetyczną średnią charakteryzującą populację generalną (testy dla pojedynczej próby), albo czy średnie reprezentujące kilka prób badanych są sobie równe. Ogólny algorytm działania tych procedur nazywa się testowaniem istotności i polega na stawianiu hipotez statystycznych oraz orzekaniu o ich prawdziwości lub fałszywości. Pod pojęciem istotności statystycznej (*statistical significance*) będziemy rozumieli prawdopodobieństwo popełnienia błędu przy weryfikacji hipotezy statystycznej, czyli ryzyko tego, że nasz werdykt nie będzie pokrywał się z rzeczywistością (*zobacz poniżej*).

Wśród testów istotności wyróżniamy testy parametryczne (to znaczy takie, które zakładają normalność rozkładu obserwacji, któremu przypisać można dwa parametry: średnią i wariancję) oraz testy nieparametryczne (które do analizy nie wykorzystują bezpośrednio wartości rejestrowanych, lecz nadawane im rangi).

W tej części zajmiemy się jedynie testami parametrycznymi; procedury nieparametryczne zostaną omówione w osobnym rozdziale.

Zależność między testami istotności i przedziałami ufności

Czy istnieje jakaś zależność między testowaniem istotności różnic a wyznaczaniem przedziałów ufności? Podczas gdy przedział ufności podaje nam zakres wartości zmiennej wokół średniej populacji (μ), w którym z określonym prawdopodobieństwem powinniśmy spodziewać się odnaleźć obserwacje z próby badanej, to test istotności określa nam czy zebrane przez nas obserwacje są spójne z jakąś wartością hipotetyczną. Jeżeli – zgodnie z testem istotności – wynik różni się od owej hipotetycznej wartości przy przyjętym poziomie istotności, to znaczy, że wynik ten nie znajdzie się w odpowiadającym przedziale ufności wokół średniej $\mu = 0$. Jeżeli zaś wynik okaże się nieistotny (nieistotnie różny od hipotetycznego), oznacza to, że mieści się on wewnątrz zakresu wyznaczonego przez ten przedział ufności.

Testy jednostronne i obustronne

Nasze pytanie o różnicę między wynikiem a wartością hipotetyczną można postawić w dwojaki sposób. Możemy zapytać, czy istnieje w ogóle jakakolwiek istotna różnica –

w górę lub w dół – nieważne czy nasz wynik będzie mniejszy albo większy od teoretycznej przyjętej *a priori* wartości, ale powinien od niej być różny. W takim przypadku pomijamy znak obliczonej wartości statystyki testu, ponieważ oczekujemy, iż w przypadku istotnej różnicy wynik ten będzie położony albo na lewym (gdy będzie mniejszy od teoretycznej wartości $\mu = 0$) albo na prawym (wtedy, gdy będzie większy od hipotetycznej wartości $\mu = 0$) krańcu rozkładu, w jego najbardziej peryferyjnych regionach nieobjętych tym obszarem pola pod krzywą, który odpowiada przyjętemu przez nas prawdopodobieństwu (np. dla wartości prawdopodobieństwa 95%, to „resztkowe" pole na peryferiach rozkładu będzie wynosiło 5%, czyli po 2.5% po każdej stronie wartości średniej umieszczonej centralnie, por. Ryc. 14 w części z zadaniami, przedstawiającą dystrybuantę i proporcje rozkładu normalnego). W takim przypadku mamy do czynienia z ***testem obustronnym*** (*two-tailed/two-sided test*) – niezależnie od tego, w którym obszarze istotności statystycznej symetrycznego rozkładu – prawym czy lewym – znajdzie się wynik, nazwiemy go wynikiem istotnie statystycznie różnym od wartości hipotetycznej. Z drugiej strony, nasze pytanie może być bardziej konkretne, kiedy na przykład pragniemy wykazać, że nasz wynik jest istotnie wyższy od przyjętej wartości hipotetycznej. W takiej sytuacji oczekujemy, że nasz istotnie różny wynik znajdzie się konkretnie w prawym obszarze istotności statystycznej. Typ testu, który weryfikuje powyższe złożenie nazywamy ***testem jednostronnym*** (*one-tailed/one-sided test*). Zauważmy, że przy tych samych warunkach prawdopodobieństwa (95%) rozmieszczenia wartości wokół średniej teoretycznej μ, szansa znalezienia się jakiegokolwiek wyniku w prawym obszarze istotności statystycznej wynosi 2.5%. Przy danej wartości statystyki testu opierającego się na rozkładzie symetrycznym – takim jak rozkład normalny lub rozkład t Studenta – prawdopodobieństwo znalezienia się wyniku poza przedziałem ufności jest dla testu jednostronnego zawsze dwukrotnie mniejsze niż dla testu obustronnego. Wynika to z metody liczenia takiego prawdopodobieństwa w oparciu o sumę logiczną (albo lewostronny albo prawostronny obszar $\Rightarrow$ prawdopodobieństwo dla lewostronnego obszaru + prawdopodobieństwo dla prawostronnego obszaru).

Wyboru testu jednostronnego lub obustronnego dokonujemy zawsze w oparciu o racjonalne przesłanki doświadczenia, a nie kierując się perspektywą wykazywania wyższych istotności różnic. Ta ostatnia możliwość jest kusząca, ale należy pamiętać, że wybór taki wiąże się także z większym ryzykiem niesłusznie odrzuconej hipotezy zerowej oraz fałszywego wnioskowania o występowaniu efektu. Na przykład, badając skuteczność leku nasennego na wydłużenie snu, zakładamy *a priori*, że ochotnicy przyjmujący ten lek będą spali dłużej niż ochotnicy otrzymujący *placebo*, a nie że osoby w obu grupach będą przesypiały różną ilość czasu. Uwzględnienie w analizie takiego badania także lewostronnego obszaru istotności (tzn. wpływu leku na skrócenie czasu snu) podważa bowiem w ogóle sens interesowania się tym lekiem jako środkiem nasennym. Naszym jedynym racjonalnym wyborem będzie więc tutaj test jednostronny. Tak jest zresztą z olbrzymią większością zastosowań testu sparowanego – *a priori* zakładamy występowanie jakiegoś ukierunkowanego efektu. Inaczej, jeżeli porównujemy określony parametr u pacjentów reprezentujących różne jednostki chorobowe: *a priori* nie zawsze możemy przewidzieć kierunek różnic.

Budowanie i weryfikacja hipotez badawczych

„Ja mogę się mylić, ty możesz mieć rację, i wspólnym wysiłkiem możemy zbliżyć się do prawdy..."

Karl Raimund Popper

Omawiając zasady formułowania hipotez badawczych, należy rozróżnić dwie kwestie. Z jednej strony, mówimy o hipotezie badawczej, która jest stwierdzeniem precyzującym istnienie jakiejś zależności, różnicy, mechanizmu funkcjonowania, prawdopodobieństwa zachodzenia procesu, itp. Jest to jakby hipotetyczny scenariusz procesu biologicznego.

Przykład: Statyny (inhibitory reduktazy HMG-CoA) wpływają na ekspresję PAI–1 w komórkach na drodze hamowania geranylacji białka *rho*.

Z drugiej strony mamy hipotezę w ujęciu statystycznym, która sprowadza się do potwierdzenia równości/nierówności matematycznej.

Przykład: Średnie masy ciała w dwóch grupach badanych są równe/są różne.

Pojedyncza hipoteza statystyczna dotyczy fragmentu hipotezy badawczej; stąd każdą koncepcję badawczą można sprowadzić do kilku/kilkunastu hipotez statystycznych – każda z nich będzie rewidowała słuszność pojedynczych porównań.

Przykład: Badając hipotezę o statynach chcemy wykazać, czy zastosowanie inhibitora geranylacji daje podobny efekt jak zastosowanie statyn; wykazanie takiego podobieństwa/analogii nie jest oczywiście żadnym dowodem na to, że mechanizm działania jest taki sam – jest jednak prostym sposobem weryfikacji, czy podążać dalej tym torem rozumowania: jeżeli stwierdzilibyśmy, że efekty są podobne, to dalej należałoby sprawdzić, czy statyny wpływają na metabolizm (obrót metaboliczny) geranylogeranylofosforanu oraz, czy geranylacja białka *rho* jest istotnie niższa po zastosowaniu statyn. Możliwe do sprecyzowania hipotezy statystyczne mogą mieć brzmienie:

- Hipoteza 1: szybkość zużywania geranylogeranylofosforanu jest taka sama/nie jest taka sama w obecności i nieobecności statyn.
- Hipoteza 2: stężenie geranylowanego białka *rho* jest/nie jest takie samo w obecności i nieobecności statyn.

Weryfikując te hipotezy mamy szansę ułożenia stwierdzeń/orzeczeń, na podstawie których da się stworzyć scenariusz działania statyn. Od właściwego sprecyzowania hipotezy statystycznej zależy to, czy będziemy mogli dowieść (z określonym prawdopodobieństwem) jej słuszności lub fałszywości. Hipotezy statystyczne zestawia się parami: hipotezie podstawowej (tzw. zerowej – *null hypothesis*) przeciwstawia się hipotezę przeciwną (alternatywną – *alternate hypothesis*) – w taki sposób, że jedna jest zaprzeczeniem drugiej. Formuła stawiania hipotez statystycznych jest ustalona – nie ma tutaj dużej dowolności, jak powinna brzmieć hipoteza zerowa, a jak hipoteza alternatywna. Wynika to z faktu, że możliwe jest jedynie odrzucenie hipotezy zerowej (z określonym prawdopodobieństwem), ale nigdy udowodnienie jej prawdziwości.

Precyzowanie hipotez polega na zestawieniu par przeciwieństw, np. stwierdzeń:

- Hipoteza zerowa (H_0) – fakt A jest prawdziwy.
- Hipoteza alternatywna (H_A) – fakt A jest fałszywy.

Umownie przyjęto, aby hipoteza zerowa zakładała niewystępowanie różnic, natomiast hipoteza alternatywna wskazuje na występowanie jednej lub wielu różnic.

Hipoteza może być odrzucona, jeżeli materiał dowodowy pozwala nam orzec z dużym prawdopodobieństwem, że hipoteza jest fałszywa. Z drugiej strony, hipoteza nie może być odrzucona, jeżeli nie mamy podstaw do jej zaprzeczenia.

Zaprzeczeniem hipotezy zerowej jest hipoteza alternatywna. Tylko jedna z nich może być prawdziwa, a wtedy druga musi być fałszywa, ponieważ obie hipotezy obejmują wszystkie możliwe warianty/możliwości. Konsekwencją niespełnienia równości $\mu_1 = \mu_2$ musi być zaakceptowanie nierówności $\mu_1 \neq \mu_2$.

Hipoteza zerowa postuluje, że $\mu_1 = \mu_2$. W rzeczywistości testujemy równość $\overline{X}_1 = \overline{X}_2$ i zakładamy, że wartości średnie dla prób badanych są reprezentatywne dla populacji generalnej.

Zasadą udowadniania prawdziwości nierówności $\mu_1 \neq \mu_2$ przy użyciu testu statystycznego jest obliczanie tzw. statystyki testu w oparciu o zebrane dane pomiarowe. Jeżeli statystyka porównania dwóch średnich jest równa zeru, to oznacza to, że dwie średnie są identyczne. Im bardziej wartość testu odbiega od wartości 0, tym większe jest prawdopodobieństwo, że średnie różnią się istotnie od siebie w sposób nieprzypadkowy. Innymi słowy, im większa jest wartość obliczonej statystyki, tym mniejsze są szanse, że hipoteza zerowa jest prawdziwa, a także, że obliczona różnica jest dziełem przypadku, a nie prawidłowością.

Skoro hipoteza zerowa jest nieprawdziwa, to znaczy, że prawdziwa jest hipoteza alternatywna. Ponieważ prawie nigdy nie znamy wartości rzeczywistych charakteryzujących miary położenia i rozproszenia dla danej zbiorowości, a jedynie dostrzegamy „poblask" rzeczywistości na podstawie analizy próby losowej, przeto o prawdziwości czy fałszywości hipotez statystycznych możemy orzekać z określonym prawdopodobieństwem mniej lub bardziej różnym od 1. Wartość tego prawdopodobieństwa precyzują dwa błędy statystyczne testowania hipotez.

Prawdopodobieństwo błędu I rodzaju (α) (*type I statistical error*), to prawdopodobieństwo błędnego odrzucenia hipotezy H_0 w przypadku, gdy jest ona prawdziwa. Błąd I rodzaju popełniamy, jeżeli mylnie odrzucamy prawdziwą hipotezę zerową.

Prawdopodobieństwo błędu II rodzaju (β) (*type II statistical error*), to prawdopodobieństwo błędnego odrzucenia hipotezy H_A w przypadku, gdy jest ona poprawna. Błąd statystyczny II rodzaju popełniamy, jeżeli nie odrzucamy hipotezy zerowej wtedy, gdy jest ona fałszywa.

	świat realny	
wynik testu	H_0 jest prawdziwa	H_0 jest fałszywa
odrzucić H_0	**błąd I rodzaju** (*prawdopodobieństwo = istotność*)	**wniosek słuszny** (*prawdopodobieństwo = moc testu*)
nie odrzucać H_0	**wniosek słuszny** (*prawdopodobieństwo = 1 – istotność*)	**błąd II rodzaju** (*prawdopodobieństwo = 1 – moc testu*)

Zauważmy, że w przypadku, gdybyśmy znali wyniki dla całej populacji generalnej, hipoteza zerowa musiałaby być albo prawdziwa, albo fałszywa z prawdopodobieństwem 100% – tym samym ryzyko popełnienia błędu I lub II rodzaju byłoby zerowe.

Prawdopodobieństwo błędu II rodzaju oraz moc testu bardzo istotnie zależy od liczebności próby oraz wielkości minimalnej różnicy, jaką badacz chce wykryć. Moc testu, to zdolność testu do wykrycia istotnej różnicy, jeżeli takowa naprawdę istnieje (zobacz też *„Metody estymacji liczebności próby"*).

Schemat popełniania błędów statystycznych nawiązuje do zasad legislacyjnych przy orzekaniu winy lub niewinności.

werdykt	**świat realny – oparty na faktach** *jest niewinny*	*jest winny*
nie jest winny		błąd II rodzaju
jest winny	błąd I rodzaju	

Jeżeli podsądny jest niewinny, a sąd orzeka jego winę, to popełniany jest błąd I rodzaju (podsądny wysłany zostaje do więzienia „*za niewinność*"). Z drugiej strony, jeżeli podsądny jest winny, a sąd orzeka jego niewinność, to popełniany jest błąd II rodzaju (sąd uwalnia winowajcę). Jeżeli sąd orzeka „*niewinny*" nie oznacza to, że podsądny nie popełnił winy. Podobnie – w testowaniu hipotez statystycznych – decyzja nie brzmi „*przyjąć hipotezę zerową*", lecz „*nie odrzucać hipotezy zerowej*" – to nie to samo. Dlatego w testowaniu statystycznym nigdy nie możemy udowodnić prawdziwości hipotezy zerowej – możemy ją jedynie odrzucić. Kiedy orzekamy, że wynik jest nieistotny statystycznie, nie znaczy to, że przyjmujemy hipotezę zerową, po prostu jej nie odrzucamy. Pamiętając, że to nie to samo, należy właściwie wyrażać i budować hipotezy statystyczne tak, aby dawały nam najbardziej wiarygodne podstawy do udowadniania hipotez naukowych.

Jak stwierdzić czy wynik jest rzeczywiście istotny? Nie można niestety uniknąć dowolności, co do tego jaki poziom istotności skłonni jesteśmy uznać jako rzeczywiście istotny. Oznacza to, że wybór poziomu istotności, powyżej którego wynik będzie odrzucany jako nieistotny jest wyborem arbitralnym. W praktyce oznacza to, że ostateczna decyzja w tym względzie zależy od wielu czynników, od tego czy wynik był przewidziany *a priori* czy też jedynie był odkryty *post hoc* (po fakcie) w wyniku analiz i porównań przeprowadzonych na określonej zbiorowości danych, od zebranego materiału doświadczalnego, jak i od tradycji panującej w danej dziedzinie badań. W wielu dziedzinach badań jako typową wartość graniczną poziomu istotności przyjmuje się $p < 0.05$. Poniżej tej wartości wynik oceniany jest jako statystycznie istotny. Zauważmy, że jest to wartość, która niesie w sobie dość duże ryzyko popełnienia błędu (5%). W badaniach biomedycznych wyniki istotne na poziomie $p < 0.01$ uważa się powszechnie jako statystycznie istotne, zaś wyniki istotne na poziomie $p < 0.005$ lub $p < 0.001$ postrzega się jako wysoce istotne. Należy jednak mieć świadomość, że tego typu klasyfikacje są niczym innym niż tylko konwencjami o dużej dozie dowolności, opartymi na doświadczeniu badawczym.

Aby postawienie hipotez nawiązywało do świata rzeczywistego, badacz powinien zadbać o właściwy dobór osobników/prób reprezentatywnych dla badanych populacji. Na przykład, badając wpływ leku na właściwości reologiczne krwi u mężczyzn ochotników w wieku 18-55 lat nie będziemy nigdy mogli wnioskować o skuteczności działania tego leku w ogólnej populacji, ponieważ nasza próba nie obejmuje kobiet, ludzi w starszym wieku, dzieci, itd.

Podsumowanie

- W ujęciu statystycznym weryfikacja hipotezy oznacza potwierdzanie równości/nierówności matematycznej.
- Każda pojedyncza hipoteza statystyczna dotyczy fragmentu hipotezy badawczej, a każdą koncepcję badawczą można rozłożyć na kilka czy kilkanaście hipotez statystycznych.
- Hipotezy statystyczne zestawia się parami w taki sposób, aby hipoteza podstawowa (tzw. zerowa) i przeciwstawna do niej hipoteza alternatywna wzajemnie się wykluczały.

- Umownie przyjmuje się, że hipoteza zerowa zakłada niewystępowanie różnic, zaś hipoteza alternatywna wskazuje na występowanie jednej lub wielu różnic.
- Hipoteza może być odrzucona jedynie z określonym prawdopodobieństwem, a nigdy z zupełną pewnością. Jeżeli nie mamy podstaw do zaprzeczenia hipotezy, to nie może być ona odrzucona, ale nie oznacza to, że jest prawdziwa.
- Im większa jest wartość obliczonej statystyki testu stosowanego do weryfikacji hipotez, tym mniejsze są szanse, że hipoteza zerowa jest prawdziwa.
- Jeżeli mylnie odrzucamy prawdziwą hipotezę zerową, to popełniamy błąd I rodzaju (błąd α), jeżeli zaś mylnie nie odrzucamy fałszywej hipotezy zerowej, to popełniamy błąd statystyczny II rodzaju (błąd β).
- Istotność wyniku testu statystycznego, to prawdopodobieństwo popełnienia błędu α, zaś prawdopodobieństwo odrzucenia fałszywej hipotezy zerowej, to moc testu.
- W testowaniu statystycznym nigdy nie możemy udowodnić prawdziwości hipotezy zerowej – możemy ją jedynie odrzucić. Orzeczenie nieistotności wyniku nie oznacza akceptacji hipotezy zerowej, lecz jedynie jej nieodrzucenie w danej sytuacji.

Testy istotności dla pojedynczej próby

Ta grupa metod obejmuje testy umożliwiające odpowiedź na pytanie, czy średnia próby badanej jest zgodna z hipotetyczną średnią charakteryzującą populację generalną. Oparte są one na statystyce rozkładu *t* Studenta lub rozkładu normalnego, a kryteria wyboru jednego z nich są takie same, jak przy obliczaniu przedziałów ufności. Najczęściej stosowanymi testami istotności dla pojedynczej próby są: test normalny, test *t* Studenta oraz sparowany test *t* Studenta.

Przykłady zastosowań tych procedur zamieszczono w *„Części II – Uzupełnienia, przykłady i zadania"*.

Sparowany test t

Sparowany test *t* Studenta (*paired t test, pair-matched Student t test*) jest szczególnym przypadkiem testu dla pojedynczej próby. Pozwala on na sprawdzenie, czy różnica wartości między parą obserwacji dokonanych na jednym obiekcie (np. w jakimś odstępie czasu) jest równa zero. Ogólnie, test służy do weryfikacji istotności wpływu określonego czynnika na zachowanie się zmiennej (np. wpływ leku na jakiś parametr krwi, środka nasennego na długość snu, związku chemotaktycznego na aktywność lub migrację komórek, zabiegu kardiochirurgicznego na reaktywność płytek krwi, itp.). Test ten weryfikuje hipotezę zerową mówiącą, że średnia różnica między wartościami dwóch zmiennych dobranych parami jest równa zeru (doświadczalne $\bar{x} = 0$). Jeżeli nie odrzucimy hipotezy zerowej, to zgadzamy się przyjąć, że wartość zmiennej w przypadku działania jakiegoś czynnika i w przypadku jego nieobecności jest taka sama (teoretyczna $m = 0$). Jeżeli natomiast stwierdzamy występowanie różnicy, to jej przyczyny mogą być dwojakie. Po pierwsze, różnice mogą wynikać ze zmienności związanej z losowaniem próby. Nawet, jeżeli czynnik rzeczywiście nie ma wpływu na wartość zmiennej, to w praktyce różnica między parami obserwacji rzadko kiedy jest dokładnie równa zeru. Druga przyczyna może być taka, że badany czynnik rzeczywiście wpływa na zmiany wartości zmiennej. Sparowany test istotności

t Studenta pozwala nam określić, które wyjaśnienie występowania niezerowych różnic jest bardziej prawdopodobne. Hipoteza zerowa testu zakłada, że pierwsze wyjaśnienie (przypadkowe występowanie niezerowych różnic) jest prawdziwe. Test pozwala nam obliczyć prawdopodobieństwo zdarzenia, że obserwowana niezerowa różnica jest dziełem przypadku wtedy, gdy hipoteza zerowa jest prawdziwa. Innymi słowy, obliczamy prawdopodobieństwo uzyskania różnicy tak dużej lub większej niż ta obliczona na podstawie obserwacji. Prawdopodobieństwo to jest naszym poziomem istotności. Jeżeli różnica między *n* parami obserwacji posiada rozkład normalny, to wartość $(\bar{x}-\mu)/(s/\sqrt{n})$ należy do pola pod krzywą rozkładu *t* Studenta o *n*–1 stopniach swobody. Ponieważ hipoteza zerowa zakłada, że teoretyczna różnica między parami obserwacji wynosi zero, statystyka sparowanego testu *t* wynosi:

$$t_{\text{par}} = \frac{\bar{x}}{s/\sqrt{n}} \qquad d.f. = n - 1$$

Obliczoną wartość $t_{\text{dośw}}$ porównujemy z wartością teoretyczną odczytaną w tablicach rozkładu *t* Studenta dla określonej liczby stopni swobody oraz przyjętego poziomu istotności (prawdopodobieństwa popełnienia błędu przy wnioskowaniu, że obserwowana różnica nie jest spowodowana przez czysty przypadek). Jeżeli $t_{\text{dośw}} \geq t_{\text{teor}}$ to z prawdopodobieństwem nie mniejszym od przyjętego (przy przyjętym poziomie istotności $<p$) możemy odrzucić hipotezę zerową i przyjąć hipotezę alternatywną; odwrotnie, jeżeli $t_{\text{dośw}} < t_{\text{teor}}$ to przy przyjętym poziomie istotności nie mamy podstaw do odrzucenia hipotezy zerowej (ale nie oznacza to, że hipoteza ta jest słuszna, *zobacz wyżej*). Innymi słowy, czym większa obliczona wartość *t* ($t_{\text{dośw}}$) (znak wyrażenia pomijamy), tym mniej prawdopodobne jest, aby niezerowa różnica między parami obserwacji była uzyskana przez czysty przypadek. Czym niższy poziom istotności tym wynik jest bardziej znamienny, gdyż tym bardziej przemawia za odrzuceniem hipotezy zerowej.

Warto uświadomić sobie, że w przypadku stosowania testu sparowanego interesujemy się właściwie wartościami różnic między parami obserwacji, nie zaś samymi wartościami obserwacji. Toteż nasze wymagania dotyczące np. normalności rozkładu zmiennej stosują się także do owych różnic, a nie do samych zmiennych. To rozkład różnic powinien być normalny, abyśmy mogli korzystać z „*udogodnień*" statystyki rozkładu normalnego. Nie jest także zasadne porównywanie wariancji w badanych grupach, gdyż parametrami rozkładu opisującego pary naszych obserwacji są średnia różnic i wariancja różnic (nie zaś obserwacji cząstkowych).

Test t Studenta dla pojedynczej próby

Test ten weryfikuje hipotezę, czy średnia próby jest istotnie różna od hipotetycznej średniej μ (lub średniej charakteryzującej populację ogólną). Zgodnie z rozważaniami na temat testu sparowanego, wartość statystyki testu *t* Studenta dla pojedynczej próby określa równanie:

$$t = \frac{\bar{x}-\mu}{s/\sqrt{n}}$$

dla liczby stopni swobody (*d.f.*) = *n* – 1.

Test normalny (test z) (*normal test, z test*)

Podobnie jak w przypadku szacowania przedziałów ufności, zamiast testu *t* Studenta wykorzystuje się zamiennie test normalny, gdy liczebność prób jest duża (przynajmniej 60 przypadków) lub w rzadko spotykanych sytuacjach, gdy znana jest wartość odchylenia standardowego populacji (σ). Wartość statystyki tego testu obliczamy według równań:

$$\text{dla dużych prób: } z = \frac{\bar{x} - \mu}{s/\sqrt{n}} \qquad \text{przy znanej wartości } s\text{: } z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}.$$

Testy istotności dla porównywania dwóch prób

Istotą testów dla porównań dwóch lub więcej grup jest badanie czy zmienność międzygrupowa (*between-group variability*) przeważa nad zmiennością wewnątrzgrupową (*within-group variability*). Jeżeli zakres zmienności obserwowanej wewnątrz każdej grupy jest mniejszy niż zakres zmienności cechy między porównywanymi zbiorowościami obserwacji, to intuicyjnie dostrzegamy, że zbiorowości takie mogą być odseparowane od siebie i tworzyć izolowane populacje. Zasada działania tych testów jest podobna jak w przypadku testów istotności dla pojedynczej próby. Równania służące do obliczania statystyki testu są jednak różne, ponieważ testy te wymagają dodatkowo, aby odchylenia standardowe porównywanych ze sobą prób były nie istotnie różnie od siebie. Porównywanie średnich przy użyciu tych testów wymaga ponadto, aby próby były niezależne. Testów tych nie powinno się zatem stosować do porównywania średnich estymowanych dla tej samej próby – w takim przypadku stosujemy omówiony wcześniej sparowany test *t* Studenta (test dla prób zależnych). Testów tych nie stosujemy także do porównywania więcej niż dwóch średnich (na zasadzie „każda z każdą") – w takich przypadkach nieodzowne jest zastosowanie poprawki Bonferroniego, lub wykorzystanie analizy wariancji i testów porównań wielokrotnych (*post hoc*).

Ogólnie, testy te pozwalają na obliczanie wartości statystyki dla określonej różnicy między dwoma próbami. Aby można się było wypowiedzieć na temat tego, w jakim regionie obszaru pod krzywą leży obliczona wartość statystyki testu, musimy wiedzieć, z jakim typem rozkładu mamy do czynienia. Podobnie jak w poprzednich rozważaniach, jest nam bardzo wygodnie przyjąć, że jest to rozkład normalny, czyli że poszczególne wartości różnic obliczanych między średnimi dla wylosowanych prób losowych z dwóch porównywanych populacji mają rozkład normalny. Warunek ten jest spełniony, jeżeli każda z porównywanych populacji ma rozkład normalny. Średnia ze wszystkich estymowanych różnic wynosi wtedy $\mu_1 - \mu_2$. Hipoteza zerowa ma wtedy taką postać, że zakłada ona równość średnich w porównywanych populacjach. Zmienność różnic opisuje równanie będące kombinacją miar rozrzutu obu populacji przy uwzględnieniu liczebności prób:

$$SE = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

Test normalny

Jest stosowany wtedy, gdy badane próby mają dużą liczebność lub w rzadkich przypadkach, gdy znamy wartości odchyleń standardowych populacji:

dla dużych prób: $z = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{(s_1^2/n_1 + s_2^2/n_2)}}$ kiedy znamy σ: $z = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{(\sigma_1^2/n_1 + \sigma_2^2/n_2)}}$.

Przedział ufności

dla dużych prób: $CI = (\bar{x}_1 - \bar{x}_2) \pm (z' \times SE)$ $SE = \sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}$

kiedy znamy σ: $CI = (\bar{x}_1 - \bar{x}_2) \pm (z' \times SE)$ $SE = \sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}$.

Test t Studenta (*Student t test for independent variables*)

Stosujemy go w przypadku prób o małej liczebności. Wymaga on, podobnie jak test istotności dla pojedynczej próby, aby rozkłady cech w porównywanych populacjach były normalne, ale jest stosunkowo mało wrażliwy na niespełnienie tego warunku. Innym warunkiem jest równość odchyleń standardowych w obu porównywanych populacjach. Błąd standardowy dla średniej różnic wynosi:

$$SE = \sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}} \quad \text{lub} \quad SE = \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

gdzie σ oznacza wspólną wartość odchylenia standardowego dla dwóch populacji o cząstkowych odchyleniach σ_1 i σ_2 z liczbą stopni swobody $(n_1 - 1) + (n_2 - 1) = n_1 + n_2 - 2$.

Na wartość takiego wspólnego odchylenia dla dwóch populacji o nierównych liczebnościach większy wpływ ma odchylenie liczniejszej próby, które jest bardziej wiarygodne:

$$s = \sqrt{\left[\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 + n_2 - 2)}\right]}.$$

Odpowiednio, statystyka testu *t* jest liczona jako:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{(1/n_1 + 1/n_2)}}, \quad d.f. = n_1 + n_2 - 2$$

zaś przedział ufności jako:

$$CI = (\bar{x}_1 - \bar{x}_2) \pm (t' \times SE), \quad SE = s\sqrt{(1/n_1 + 1/n_2)}.$$

W przypadku heteroscedastyczności (nierówność wariancji) pożądane jest najpierw spróbować poddać dane transformacji tak, aby można było zastosować test *t* Studenta (zobacz *„Transformacja danych i odstające obserwacje – sposoby „normalizacji" rozkładu"*). Alternatywnie, można wykorzystać testy nieparametryczne (np. test U Manna-Whitneya, test sumy rang Wilcoxona).

Podsumowanie

Test *t* Studenta do porównywania dwóch prób zakłada, że obie próby zostały wylosowane z populacji o rozkładzie normalnym o równych wariancjach. Nie zawsze jednak można być pewnym, że oba te kryteria są spełnione. Szczęśliwie dla badacza, test *t* jest bardzo „odporny" na naruszenie tych założeń, szczególnie w przypadkach, gdy próby są jednakowo lub prawie jednakowo liczne, oraz gdy stosujemy test obustronny. Dobrze jest pamiętać, że „odporność" taka wzrasta ze wzrostem liczebności próby. Jeżeli rozkłady badanych prób są wyraźnie skośne (asymetryczne), to oczywiście są to szczególne okoliczności, kiedy powinniśmy unikać testów jednostronnych – zwłaszcza testowania przy niskich poziomach istotności. Moc testu jest słabo wrażliwa na skośność w przypadku testów obustronnych, ale z kolei bardzo wrażliwa w testach jednostronnych. W przypadku nierównych wariancji obu prób, prawdopodobieństwo popełnienia błędu I rodzaju jest większe niż założona wartość α, ale jeżeli próby są w przybliżeniu jednakowo liczne, test *t* pozostaje słabo wrażliwy na nierówność wariancji dla licznych i bardzo licznych prób. Jeżeli $n_1 \neq n_2$, to prawdopodobieństwo popełnienia błędu I rodzaju jest niższe od α, gdy większa wariancja dotyczy bardziej licznej próby, i odwrotnie. Różnice wariancji do 10% są do przyjęcia, czym większa heteroscedastyczność tym większe odstępstwa rzeczywistej wartości prawdopodobieństwa od przyjętego α. Niekiedy zaleca się sprawdzenie równości wariancji przed zastosowaniem testu *t*, ale zważywszy, że test *t* jest stosunkowo słabo wrażliwy na naruszenie założenia jednorodności wariancji, oraz że testy do jej weryfikacji nie sprawdzają się w przypadku odstępstw od normalności rozkładu, częstą praktyka jest poniechanie tego etapu analizy. W przypadku poważnych naruszeń założeń stosowania testów parametrycznych oraz niewielkiej liczebności prób, korzystamy z metod nieparametrycznych.

Co powinniśmy zrobić przed przystąpieniem do testowania istotności testami t?

- Obejrzeć jak wygląda rozkład wyników – graficzna prezentacja rozkładu wyników daje nam dużo informacji nawet przed przeprowadzeniem testów normalności czy jednorodności wariancji.
- Możemy skorzystać z metod transformacji danych, co wielokrotnie zaowocuje normalizacją transformowanych danych pomiarowych.
- Nie musimy się natomiast martwić równością wariancji w przypadku testu sparowanego, gdyż to co badamy przy porównywaniu zmiennych skojarzonych to zmienność różnic par wyników.

Czego bezwzględnie nie możemy robić?

- Stosować testu *t* do badania cech o rozkładach dyskretnych.
- Stosować testów porównań dwóch prób do porównań wielokrotnych.
- Stosować testu *t* w przypadkach interakcji zmiennych kategoryzujących (czynników).

Rozdział 5

Testy istotności dla porównywania więcej niż dwóch prób – analiza wariancji

Istotą analizy wariancji, podobnie jak testów dla porównań dwóch grup, jest badanie, czy zmienność międzygrupowa przeważa nad zmiennością wewnątrzgrupową. Jeżeli zakres zmienności obserwowanej wewnątrz każdej grupy jest mniejszy niż zakres zmienności cechy między porównywanymi zbiorowościami obserwacji, to intuicyjnie dostrzegamy, że zbiorowości takie mogą być odseparowane od siebie i tworzyć izolowane populacje. Wyjaśnia to jednocześnie nazwę tej analizy – chociaż chcemy dowiedzieć się o różnicach między średnimi, testujemy tak naprawdę wariancje (zmienności) badanych cech. Zmienne, które mierzymy (czyli nasze wyniki doświadczenia), są nazywane zmiennymi zależnymi, zaś te, którymi manipulujemy, które kontrolujemy lub te, które służą nam do kategoryzacji (grupowania) obserwacji na grupy porównywane są nazywane czynnikami lub zmiennymi niezależnymi.

Celem analizy wariancji (*analysis of variance,* ANOVA) jest zazwyczaj testowanie istotności różnic pomiędzy średnimi więcej niż dwóch grup na drodze analizowania wariancji. W przypadku porównywania dwóch średnich ANOVA daje takie same wyniki, jak test *t* Studenta dla prób niezależnych.

Ogólnie, w metodzie ANOVA wykorzystuje się fakt, że wariancje mogą być rozdzielane na ich składowe komponenty. Pamiętamy, że wariancja w sensie rachunkowym jest funkcją sum kwadratów odchyleń poszczególnych wartości zmiennej od wartości ogólnej średniej (*sum of squares*, *SS*). Zmienność w obrębie każdej grupy (wyrażona jako $SS_{\text{błędu}}$) jest nazywana wariancją błędu (*within-group mean square error*) i wyrażana jako tzw. losowy błąd średniokwadratowy ($MS_{\text{błędu}}$). Jest ona rachunkowo „*nierozkładalna*" na komponenty składowe i dlatego jest często nazywana zmiennością niewyjaśnioną (*unexplained variability*); odpowiada ona błędowi losowemu i można ją przypisać zmienności cechy w badanej populacji. Z drugiej strony, wariancja efektu (SS_{efektu}) (*between-group mean square error*), wyrażana jako tzw. błąd średniokwadratowy efektu (MS_{efektu}) (*mean square error*), wynika zarówno ze zmienności cechy w każdej z grup, jak i z różnic średnich pomiędzy porównywanymi grupami. Wyjaśnia ona przynależność do grupy, ponieważ to ona warunkuje różnice pomiędzy średnimi. Ten rodzaj zmienności nazywamy zmiennością wyjaśnioną (*explained variability*), dlatego że całkowitą wariancję można rozłożyć (podzielić) na składową odpowiadającą błędowi losowemu (wariancja wewnątrzgrupowa, $SS_{\text{błędu}}$ w obrębie każdej z grup) oraz składowe, które odnoszą się do różnic pomiędzy średnimi (wariancja międzygrupowa, SS_{efektu} w obrębie całej zbiorowości wyników pochodzących od wszystkich grup).

Pod względem rachunkowym w metodzie analizy wariancji opieramy się na porównaniu wariancji odnoszącej się do zmienności pomiędzy grupami (MS_{efektu}) ze zmiennością w obrębie grup ($MS_{błędu}$). Innymi słowy, naszym celem w tej metodzie jest sprawdzenie, czy wariancja obliczona w oparciu o zmienność wewnątrzgrupową (tzn. dla każdej z porównywanych grup oddzielnie) jest istotnie mniejsza od wariancji oszacowanej w oparciu o całkowitą zmienność (to znaczy dla wszystkich wyników wszystkich porównywanych grup). Mniejsza wariancja wewnątrzgrupowa oraz większa globalna, to oczywiście dowód dla nas, że wyniki w poszczególnych grupach różnią się istotnie, czyli że liczone na podstawie tych różniących się wyników średnie także będą różne w różnych grupach.

Nasza hipoteza zerowa będzie zakładać, że nie ma różnic wartości średnich pomiędzy grupami w populacji. Ponieważ każda cecha charakteryzuje się jakąś zmiennością w populacji, powinniśmy nadal oczekiwać nieznacznych losowych wahań średnich dla różnych prób, na przykład w przypadku pobierania mało licznych prób. Abyśmy pozostawali w zgodzie z hipotezą zerową, wariancja estymowana w oparciu o zmienność w obrębie poszczególnych grup powinna być taka sama (nieistotnie różna), jak wariancja opisująca zmienność pomiędzy grupami. Do porównania tych dwóch oszacowań wariancji służy test *F*, oceniający, czy iloraz obu rodzajów wariancji (międzygrupowej do wewnątrzgrupowej) jest istotnie większy od 1. Wartość krytyczna *F* tego testu mówi nam *de facto,* ile razy wariancja wyjaśniona (czyli ta wynikająca z różnic między średnimi) przewyższa wariancję niewyjaśnioną (czyli tę pochodzącą od błędu losowego, opisującą zmienność wewnątrz grup). Czym wyższa jest przewaga wariancji wyjaśnionej (dotyczącej zmienności międzygrupowej) nad wariancją niewyjaśnioną (dotyczącej zmienności wewnątrzgrupowej), czyli czym wyższa jest wartość tego ilorazu, tym wyższa istotność różnic między grupami.

W pewnych przypadkach statystyka *F* może być jednak myląca, a mianowicie wtedy, gdy średnie i wariancje w obrębie grup są ze sobą skorelowane. Taki przypadek – wysoka średnia oraz duża wariancja w grupie – pojawia się często w sytuacji, gdy w obrębie danych występują odstające obserwacje. Jeden lub dwa skrajne przypadki w grupie liczącej niewiele obserwacji mogą znacznie obciążyć średnią, a także znacznie zwiększyć wariancję.

Wielowymiarowym odpowiednikiem testu *F* jest test statystyka lambda Wilksa (*Wilk's lambda statistics*).

Najprostszym rodzajem analizy wariancji jest ***jednoczynnikowa analiza wariancji*** (*one-way analysis of variance*), kiedy mamy do czynienia z pojedynczą zmienną grupującą (niezależną). W istocie, charakterystyka jakiegoś procesu biologicznego rzadko kiedy zależy od pojedynczej zmiennej i na jej zmienność wyjaśnianą maja wpływ liczne czynniki. Przykładowo, próbując wyjaśnić, w jaki sposób wzrasta ryzyko miażdżycy naczyń, powinniśmy wziąć pod uwagę wiele czynników, o których wiemy, że wpływają na parametry koagulologiczne, reologiczne, metaboliczne, biochemiczne czy strukturalne – dotyczące samej budowy ściany naczyniowej. Nawet w przypadku, gdybyśmy badali i porównywali dwie grupy pacjentów, to i tak zastosowanie zwykłego testu *t* Studenta do poszczególnych cząstkowych porównań spłycałoby nasz problem, gdyż nie dostarczałyby nam wiedzy na temat interakcji różnych czynników. W takich przypadkach uciekamy się do stosowania wieloczynnikowej analizy wariancji, tzn. takiego jej modelu, gdzie występuje więcej niż jedna zmienna grupująca (wyjaśniająca). Dla takich wieloczynnikowych modeli ANOVA obliczenia stają się coraz bardziej złożone i trudno je przeprowadzać na dużych macierzach danych (wiele zmiennych z dużymi liczebnościami) bez korzystania z pomocy statystycznych programów komputerowych. Z tego względu, a także z uwagi na ograniczoną obszerność tego opracowania, bliżej zostaną omówione jedynie jednoczynnikowa i ***dwu-***

czynnikowa analiza wariancji (*two-way analysis of variance*). Kryterium rozstrzygającym o roli czynnika (zmiennej grupującej) może być albo fakt, że zmienna określa grupę, do której należą obserwacje (zdrowi-chorzy, leczeni-nieleczeni, występowanie-niewystępowanie czynnika, itd.), albo też stanowi zmienności, które należy uwzględnić.

Na przykład, z obserwacji klinicznych wynika, że częstość występowania choroby wieńcowej jest różna w różnych grupach etnicznych. Zaplanowano badania, które mają wykazać, czy taki rozkład częstości jest spójny z częstością występowania hiperlipidemii w różnych grupach etnicznych. Ale profil lipidowy osocza zależy silnie od wieku oraz płci pacjenta, toteż czynniki te należałoby uwzględnić w analizie wyników badania, chociaż to nie ich wpływ jest dla nas interesujący w tym badaniu. Ich włączenie ma dwie korzyści. Po pierwsze, test istotności badania różnic etnicznych posiada większą moc, gdyż jest bardziej prawdopodobne, że wykryjemy przy jego użyciu jakiekolwiek naprawdę istniejące różnice. Po drugie, nie ryzykujemy, że wykryte różnice nie będą „zafałszowane" przez wpływ wieku czy płci.

Wśród czynników wyróżniamy dwa rodzaje: bardziej powszechne czynniki stałe (tzw. *fixed effects*), których występowanie jest immanentnie przypisane do przedmiotu badań (np. płeć, grupa wiekowa, polimorfizm genetyczny) i które przyjmują zawsze konkretną wartość dla danego obiektu badań (np. kobieta-mężczyzna, starszy-młodszy) (jest to tzw. model I analizy wariancji, ANOVA 1), oraz czynniki losowe (tzw. *random effects*), których źródłem jest błąd losowy (np. dokładność wykonania czegoś przez różne osoby/grupy osób, konkretny wykonawca jest tutaj czynnikiem losowym) (w modelu II analizy wariancji, ANOVA 2). Z uwagi na występowanie więcej niż jednego czynnika (kilku zmiennych niezależnych) czynniki te modyfikują wzajemnie swoje wpływy na zmienne zależne, tzn. dochodzi do interakcji między nimi. Możemy ogólnie stwierdzić, że w takiej sytuacji jeden efekt jest modyfikowany (warunkowany) przez inny efekt. W przypadku najprostszej sytuacji, gdy występuje interakcja pomiędzy dwoma czynnikami, główny efekt (np. występowanie choroby) jest modyfikowany przez drugi czynnik (np. palenie tytoniu). W przypadku trójczynnikowej interakcji możemy stwierdzić, że dwuczynnikowa interakcja pomiędzy grupą etniczną i płcią jest modyfikowana (warunkowana) przez trzeci czynnik wiek. W miarę zwiększania liczby czynników (zmiennych grupujących) sytuacja robi się coraz bardziej złożona. Na przykład, mając do czynienia z czteroczynnikową interakcją, możemy określić, że trójczynnikowa interakcja jest modyfikowana poprzez wpływ czwartej zmiennej, a co więcej, mogą istnieć różne rodzaje interakcji na różnych poziomach oddziaływania czwartej zmiennej. Kiedy złożoność takich interakcji wzrasta jeszcze bardziej, korzystamy niekiedy z bardziej ogólnych technik, takich jak na przykład regresja wielokrotna, która w aspekcie rachunkowym jest o wiele bardziej złożoną techniką. W prostych sytuacjach, kiedy czynników nie jest wiele, metoda analizy wariancji skuteczniej charakteryzuje nasz układ badany i dlatego korzystamy z niej częściej.

Założenia i warunki stosowania analizy wariancji oraz konsekwencje ich naruszenia

Dwa podstawowe założenia stosowania testu *F,* to normalność rozkładu oraz jednorodność wariancji. Parametryczna analiza wariancji wymaga ponadto, aby zmienne zależne miały charakter ciągły i były wyrażone przynajmniej na skali przedziałowej. Test *F* jest w znacznym stopniu odporny na odchylenia od normalności, o ile rozkład nie jest bardzo leptokurtyczny lub platykurtyczny. Jeżeli kurtoza jest wyraźnie większa od 0, to wartość *F* zbliża się do małych wartości, czyli nie możemy odrzucić hipotezy zerowej, nawet wtedy, gdy nie jest prawdziwa. Przeciwnie, gdy wartość kurtozy jest mniejsza od 0 wzrasta ryzyko odrzucenia hipotezy zerowej nawet wtedy, gdy nie jest ona fałszywa. Mały wpływ na wartość statystyki *F* ma skośność rozkładu. Przy dużych liczebnościach – podobnie jak to ma miejsce w przypadku rozkładu *t* – odchylenia od rozkładu normalnego nie mają w ogóle znaczenia z uwagi na centralne twierdzenie graniczne.

Naruszenie założenia o jednorodności wariancji mogłoby mieć poważniejsze konsekwencje. Gdy wariancje w porównywanych grupach różnią się znacznie między sobą, wówczas ich sumowanie w celu oszacowania *„wspólnej"* wariancji wewnątrzgrupowej nie jest właściwe, ponieważ nie istnieje wtedy żadna *„wspólna"* wariancja. W takim przypadku może pomóc transformacja danych. Z praktyki wynika, że założenie to może mieć poważne znaczenie jedynie przy małych liczebnościach porównywanych grup.

Wyniki różnych badań i analiz pokazały, że tylko w przypadku bardzo poważnych naruszeń tych założeń musimy zainteresować się elastycznością statystyki *F*. Ogólnie, statystyka *F* (w ANOVA) jest uważana za silny test istnienia różnic między średnimi, przy założeniach że:

- liczebności n_i w grupach są większe niż 10, oraz
- średnie nie są skorelowane z odchyleniami standardowymi w grupach.

Jednoczynnikowa analiza wariancji

Jest to najprostsza wersja ANOVA, którą wykorzystujemy do prostego porównywania cechy w kilku grupach badanych. Analiza ta opiera się na szacowaniu jaką porcję całkowitej wariancji stanowi wariancja, którą możemy przypisać różnicom między grupami. W rachunku obliczeniowym całkowita suma kwadratów różnic od średniej (*SS*) jest rozdzielana na dwie składowe (komponenty):

- *SS* pochodzącą od różnic między grupami oraz
- *SS* pochodzącą od różnic między poszczególnymi obserwacjami w każdej z grup (jest to tzw. resztowa suma kwadratów) (*residual sum of squares*).

Na podstawie obliczonych sum kwadratów obliczamy błędy średniokwadratowe, czyli porcje poszczególnych sum kwadratów przypadające na pojedynczy stopień swobody dla porównań wewnątrzgrupowych i międzygrupowych.

Wartość statystyki testu *F* (testu ilorazu wariancji) obliczamy jako:

$$F = \frac{MS_{międzygrupowa}}{MS_{wewnątrzgrupowa}} = \frac{MS_{efektu}}{MS_{błędu}},$$

zaś liczbę stopni swobody jako:

$$d.f. = d.f._{\text{międzygrupowe}},\ d.f._{\text{wewnątrzgrupowe}} = (k–1,\ N–k).$$

F powinno być bliskie 1, jeżeli nie ma różnic między grupami, gdyż wtedy zmienność wewnątrz każdej z grup równa jest zmienności wewnątrz grupy „*wspólnej*", tzn. powstałej ze scalenia wszystkich grup w jedną. Jeżeli natomiast *F* jest większe od 1, to znaczy, że zmienność wewnątrz grupy „wspólnej" przewyższa zmienność wewnątrz każdej z grup. Jest to możliwe tylko wtedy, gdy cząstkowe obserwacje w jednej z grup mają wyraźnie wyższe lub niższe wartości niż w innych grupach. W zgodzie z założeniem hipotezy zerowej, że obserwowane różnice są wynikiem czystego przypadku, wartości obliczonej statystyki *F* mają rozkład *F*. Rozkład ten tym się różni od innych rozkładów, że definiowany jest dla pary stopni swobody: *k*–1 stopni swobody w liczniku, oraz *N*–*k* stopni swobody w mianowniku. Jak w przypadku wszystkich testów istotności, obliczone wartości *F* porównujemy z tablicowymi: dla przypadków, gdy $F_{\text{dośw}} > F_{\alpha,\ k–1,\ N–k}$, mamy podstawy, aby z prawdopodobieństwem α odrzucić hipotezę zerową*.

Jednoczynnikowa analiza wariancji jest rozwinięciem testu *t* Studenta. W przypadku gdy mamy do czynienia z porównywaniem jedynie dwóch grup, wyniki obu testów są identyczne. Wartość statystyki *F* równa jest liczbowo wartości *t* podniesionej do kwadratu, a wartości krytyczne testu *F* przy (1, *N*–2) stopniach swobody są identyczne z wartościami t^2 przy *N*–2 stopniach swobody.

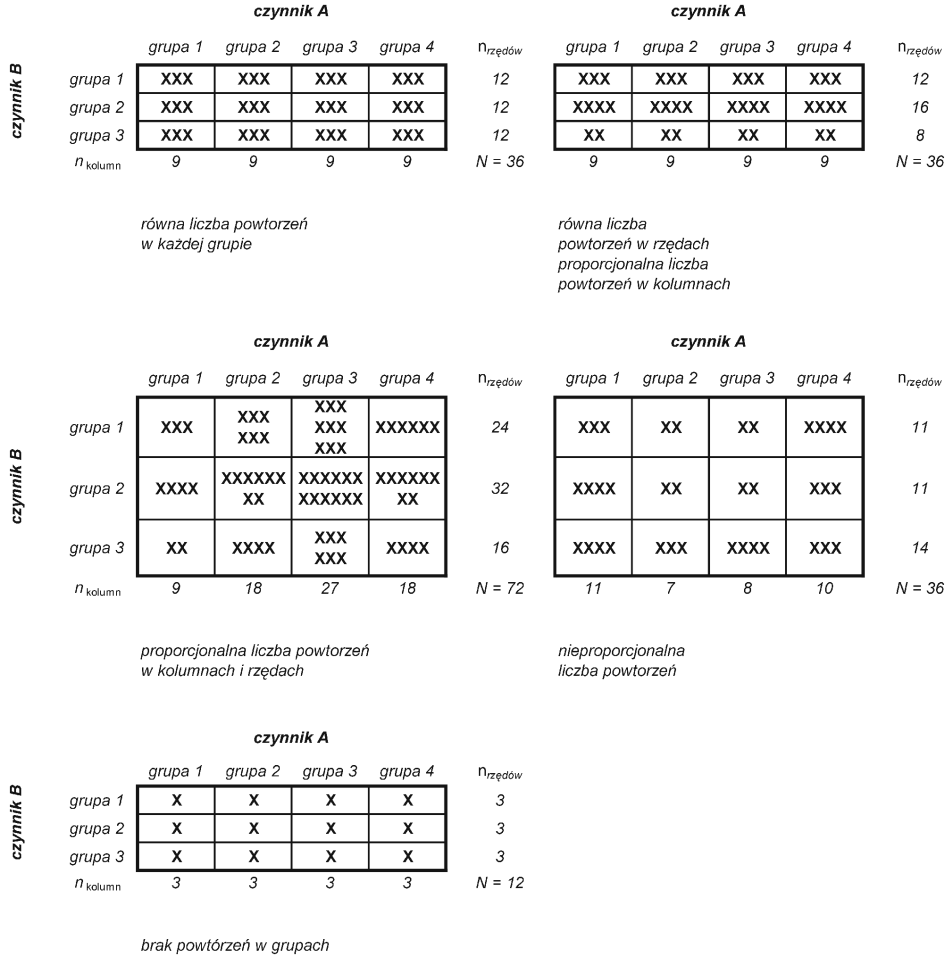
Dwuczynnikowa analiza wariancji

Ten typ analizy wariancji stosujemy w przypadkach gdy chcemy uwzględnić wpływ dwóch różnych czynników na zmienność obserwacji. Na przykład, badając liczbę płytek krwi możemy pogrupować dane ze względu na płeć i grupę wiekową. Oba czynniki mogą podlegać wzajemnym wpływom – mówimy wtedy o interakcjach między czynnikami i uwzględniamy ich wymiar w interpretacji wyników (Ryc. 6). Jeżeli w każdej wydzielonej grupie wystepuje jednakowa liczba obserwacji, stosujemy wariant dwuczynnikowej analizy wariancji z równą liczebnością grup (*balanced design*), w przeciwnym wypadku wybieramy wariant metody zwany analizą wariancji z różną liczebnością grup (*unbalanced design*). W praktyce spotykamy się z koniecznością zestawiania obserwacji w zgodzie z jednym z dwóch modeli: pomiary z powtórzeniami (z więcej niż jedną obserwacją w grupie) lub pomiary bez powtórzeń (pojedyncze obserwacje na grupę).

Kiedy niektóre z czynników pojawiają się w obrębie poziomów innego czynnika, mamy do czynienia z ***układem hierarchicznym analizy wariancji*** (*hierarchical analysis of variance*) (tzw. zagnieżdżony model analizy wariancji – *nested design*).

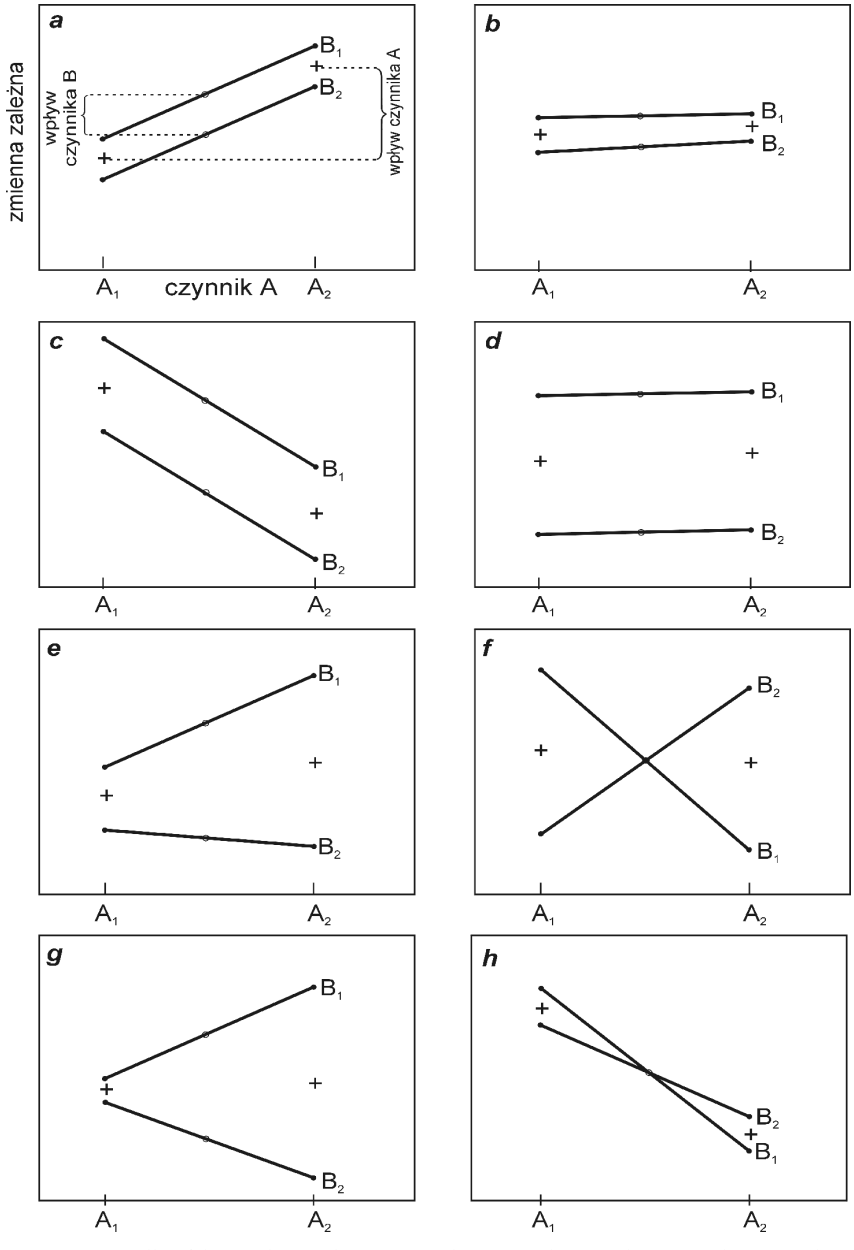
Szczegóły obliczeń w zakresie analizy danych w każdym z takich modeli podano w *„Części II – Uzupełnienia, przykłady i zadania"*.

* Wartości krytyczne rozkładu *F* są z założenia zawsze jednostronne, gdyż wartość *F* jest z założenia zawsze większa od 1.



Układy z czynnikami międzygrupowymi oraz czynnikami powtarzanych pomiarów (analysis of variance for repeated measures)

Wspomniano wcześniej, że chcąc porównać dwie grupy należałoby zastosować (pod warunkiem spełnienia założeń testów parametrycznych) test *t* dla prób niezależnych, chcąc natomiast porównać dwie zmienne dotyczące tych samych osobników (obserwacji), zastosowalibyśmy test *t* dla prób zależnych. Takie rozróżnienie – dotyczące grup zależnych i niezależnych – jest ważne również w przypadku analizy wariancji, gdy mamy do czynienia z powtarzanymi pomiarami dla tej samej zmiennej (przy różnych warunkach pomiaru lub w różnych momentach czasowych) dla tych samych grup. Wówczas jeden z czynników (zmiennych) grupujących jest czynnikiem powtarzanych pomiarów (nazywamy go czynnikiem wewnątrzobiektowym (*within-group factors*), ponieważ w celu oszacowania jego istotności obliczamy wewnątrzgrupowe sumy kwadratów; jest to zmienność losowa w obrębie grupy). Jeśli porównujemy różne grupy badanych (np. kobiety i mężczyzn, różne jednostki chorobowe, itd.), wówczas traktujemy kontrolowany czynnik jako czynnik międzygrupowy (*between-group factors*).



Rys. 6. Różne przypadki efektów głównych oraz interakcji czynników w dwuczynnikowej analizie wariancji; miarą wpływu czynnika A jest stopień nachylenia linii łączącej punkty oznaczone (+), miarą wpływu czynnika B jest odległość między odcinkami oznaczonymi B_1 i B_2, miarą interakcji między czynnikami jest różnica w stopniu nachylenia prostych (porównanie nachyleń prostych zobacz też str. 112): a) duży wpływ czynnika A, mały wpływ czynnika B, brak interakcji; b) brak wpływu czynnika A, słaby wpływ czynnika B, brak interakcji; c) duży wpływ czynnika A, duży wpływ czynnika B, brak interakcji; d) brak wpływu czynnika A, bardzo duży wpływ czynnika B, brak interakcji; e) słaby wpływ czynnika A, duży wpływ czynnika B, silna interakcja; f) brak wpływu czynnika A, brak wpływu czynnika B, interakcja między A i B; g) brak wpływu czynnika A, duży wpływ czynnika B, silna interakcja; h) duży wpływ czynnika A, brak wpływu czynnika B, nieznaczna interakcja (wg Zara, 1999).

W wielu przypadkach nasz model doświadczalny wymaga uwzględnienia zarówno czynników międzygrupowych, jak i czynników powtarzanych pomiarów. Badając na przykład stężenie PAI–1 wśród pacjentów z chorobą niedokrwienną serca oraz ludzi zdrowych przychodzących na okresowe badania lekarskie (grupa kontrolna i grupa pacjentów to czynnik międzygrupowy), obok efektu głównego mamy do czynienia ze zmiennością w różnych punktach czasowych w każdej z badanych grup. Kilka pomiarów przeprowadzanych u tych samych pacjentów podczas kolejnych niezależnych wizyt lekarskich (odbywających się w różnych momentach czasowych) stanowi czynnik powtarzanych pomiarów. Sposób interpretacji efektów głównych i interakcji będzie wynikiem tego, jak oba czynniki mogą wzajemnie oddziaływać na siebie (np. pacjenci z chorobą niedokrwienną serca mogą być bardziej wrażliwi na częste wahania ciśnienia atmosferycznego).

Niekiedy mamy do czynienia z przypadkami, kiedy możemy pominąć efekty interakcji. Sytuacja taka występuje dość często w praktyce wówczas, gdy na przykład 1) nie możemy przeprowadzić pełnego układu doświadczalnego z przyczyn ekonomicznych, lub też gdy 2) wiemy, że w danej populacji efekt interakcji jest na tyle nieistotny, że możemy go pominąć. Na przykład planujemy przeprowadzić badanie, w którym chcemy sprawdzić skuteczność 4 różnych antagonistów receptora dla fibrynogenu w hamowaniu agregacji płytek krwi. Do badania wykorzystamy krew uzyskaną od 4 zdrowych dawców płci męskiej w wieku od 18 do 23 lat. Badania należy przeprowadzić jak najszybciej po pobraniu krwi, toteż badania wszystkich antagonistów nie jest w stanie przeprowadzić w tym samym czasie jeden laborant. Nasz plan zakłada, że każdy z 4 wyznaczonych specjalistów techników laboratoryjnych będzie w danym czasie badał wpływ jednego antagonisty, wykorzystując do tego krew pobraną od jednego pacjenta.

Pełne doświadczenie zestawione według takiego hierarchicznego układu z uwzględnieniem wszystkich czynników (rodzaj antagonisty, dawca krwi, laborant), tzn. takie, w którym każda kombinacja laboranta, antagonisty oraz dawcy krwi pojawia się przynajmniej jeden raz wymagałaby $4 \times 4 \times 4 = 64$ grup. Jednakże możemy nie mieć środków ani czasu, aby przeprowadzić próby we wszystkich kombinacjach, a ponadto, wydaje się mało prawdopodobne, aby np. osoba laboranta występowała w interakcji z dawcą krwi lub rodzajem antagonisty w stopniu, który mógłby mieć jakieś racjonalne praktyczne znaczenie. Biorąc to pod uwagę moglibyśmy w rzeczywistości zrealizować jedynie tzw. układ kwadratu łacińskiego (*analysis of variance of Latin squares*) obejmującego 4 rodzaje antagonistów (A, B, C i D) i 16 osobnych grup badanych.

	dawca krwi			
	1	2	3	4
laborant 1	A	B	C	D
laborant 2	B	C	D	A
laborant 3	C	D	A	B
laborant 3	D	A	B	C

Widzimy, że układ ten jest układem hierarchicznym niekompletnym w tym sensie, że nie wszystkie kombinacje grup dla poszczególnych czynników są uwzględnione w modelu. Na przykład, laborant 1 będzie badał płytki krwi od dawcy 1 z dodatkiem antagonisty A, podczas gdy laborant 3 będzie badał krew od tego samego dawcy z dodatkiem antagonisty C. Co więcej, poszczególne grupy czynnika, rodzaj antagonisty (A, B, C i D) są rozmieszczane w sposób przypadkowy w macierzy wyznaczonej przez czynniki – dawca krwi i laborant. Podobne rozwiązania są jednak bardzo często stosowane w praktyce planowania badań w sytuacjach, gdy niektóre efekty interakcji możemy pominąć bez szkody dla wyników analizy.

Dwuczynnikowa analiza wariancji z równą liczbą powtórzeń w grupach *(balanced design with replication)*

W metodzie tej obserwacje są pogrupowane ze względu na dwa różne czynniki, które mogą wchodzić w interakcje między sobą, przy czym każda grupa zawiera więcej niż jedną obserwację i liczba tych obserwacji (powtórzeń) jest jednakowa we wszystkich grupach. Naszym zadaniem jest stwierdzenie, na ile jeden czynnik i na ile drugi czynnik wpływają na wartość obserwacji. Zmienność w każdej z grup wynika z powtarzalności obserwacji (zmienność losowa), lub zmienności międzyosobniczej. Całkowita wariancja może być rozdzielona na cztery składniki:

- sumę kwadratów efektu głównego w zakresie różnic między grupami czynnika pierwszego,
- sumę kwadratów efektu głównego w zakresie różnic między grupami czynnika drugiego,
- sumę kwadratów różnic w zakresie interakcji między pierwszym i drugim czynnikiem,
- resztową sumę kwadratów różnic między obserwacjami w każdej grupie.

Poszczególne sumy kwadratów dla modelu ANOVA 1 (czynniki stałe) obliczamy w następujący sposób:

zmienność	**suma kwadratów (*SS*)**	**stopnie swobody (*d.f.*)**	**błąd średnio-kwadratowy (*MS*)**
całkowita	$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{l=1}^{n} X_{ijl}^2 - C$	N–1	
grupowa	$\frac{\sum_{i=1}^{a}\sum_{j=1}^{b}\left(\sum_{l=1}^{n} X_{ijl}\right)^2}{n} - C'$	ab–1	
czynnika 1 (A)	$\frac{\sum_{i=1}^{a}\left(\sum_{j=1}^{b}\sum_{l=1}^{n} X_{ijl}\right)^2}{bn} - C'$	a–1	$\frac{SS_A}{d.f._A}$
czynnika 2 (B)	$\frac{\sum_{j=1}^{b}\left(\sum_{i=1}^{a}\sum_{l=1}^{n} X_{ijl}\right)^2}{an} - C$	b–1	$\frac{SS_B}{d.f._B}$
interakcja A x B	$SS_{\text{grup}} - SS_{\text{A}} - SS_{\text{B}}$	(a–1)(b–1)	$\frac{SS_{A\times B}}{d.f._{A\times B}}$
błędu (*wewnątrzgrupowa*)	$SS_{\text{całk}} - SS_{\text{grup}}$	$ab(n-1)$= $d.f._{\text{całk.}} - d.f._{\text{grup}}$	$\frac{SS_{błędu}}{d.f._{błędu}}$

a jest liczbą grup czynnika A, podobnie *b* jest liczbą grup czynnika B, zaś *n* jest liczbą powtórzeń w każdej grupie.

Wartość C obliczamy z równania:

$$C = \frac{\left(\sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{l=1}^{n} X_{ijl} \right)^2}{N},$$

gdzie $N = abn$.

Występowanie istotnej interakcji oznacza, że wpływ czynnika B na czynnik A i odwrotnie nie jest jednakowy we wszystkich grupach czynnika A i/lub B. W sytuacjach takich, nawet jeżeli stwierdzimy istotność różnic pomiędzy grupami w zakresie każdego z czynników, nie ma ona praktycznego znaczenia i jest trudna do interpretacji.

W przypadku modelu ANOVA 2 (czynniki losowe) wartości statystyki F liczymy odmiennie:

badany efekt	**model 1 (oba czynniki: A i B stałe)**	**model 2 (oba czynniki: A i B losowe)**	**model 3 (czynnik A stały: czynnik B losowy)**
czynnik A	$\frac{MS_A}{MS_{błędu}}$	$\frac{MS_A}{MS_{A\times B}}$	$\frac{MS_A}{MS_{A\times B}}$
czynnik B	$\frac{MS_A}{MS_{błędu}}$	$\frac{MS_B}{MS_{A\times B}}$	$\frac{MS_B}{MS_{błędu}}$
interakcja A x B	$\frac{MS_{A\times B}}{MS_{błędu}}$	$\frac{MS_{A\times B}}{MS_{błędu}}$	$\frac{MS_{A\times B}}{MS_{błędu}}$

Interpretację wartości statystyki F powyższych składowych znajdzie Czytelnik w *„Części II – Uzupełnienia, przykłady i zadania"* niniejszego opracowania.

Dwuczynnikowa analiza wariancji z nierówną liczbą powtórzeń w grupach *(analysis of variance, unbalanced design)*

Ten bardziej złożony od poprzedniego model występuje bardzo często w praktyce badawczej, gdy nie jesteśmy w stanie zagwarantować jednakowej liczebności grup. Zakłada on, że liczebności w każdej grupie *ij* dla *i*-tej grupy czynnika A i *j*-tej grupy czynnika B są proporcjonalne, tzn.:

$$n_{ij} = \frac{(n_i)(n_j)}{N}.$$

Sumy kwadratów dla tego modelu dwuczynnikowej analizy wariancji obliczamy w następujący sposób:

zmienność	**suma kwadratów (*SS*)**	**stopnie swobody (*d.f.*)**	**błąd średnio-kwadratowy (*MS*)**
całkowita	$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{l=1}^{n_{ij}} X_{ijl}^2 - C$	N–1	
grupowa	$\sum_{i=1}^{a}\sum_{j=1}^{b}\frac{\left(\sum_{l=1}^{n_{ij}} X_{ijl}\right)^2}{n_{ij}} - C$	ab–1	$\frac{SS_{grup}}{d.f._{grup}}$
czynnika 1 (A)	$\sum_{i=1}^{a}\frac{\left(\sum_{j=1}^{b}\sum_{l=1}^{n_{ij}} X_{ijl}\right)^2}{\sum_{j=1}^{b} n_{ij}} - C$	a–1	$\frac{SS_A}{d.f._A}$
czynnika 2 (B)	$\sum_{j=1}^{b}\frac{\left(\sum_{i=1}^{a}\sum_{l=1}^{n_{ij}} X_{ijl}\right)^2}{\sum_{i=1}^{a} n_{ij}} - C$	b–1	$\frac{SS_B}{d.f._B}$
interakcja A x B	$SS_{\text{grup}} - SS_{\text{A}} - SS_{\text{B}}$	(a–1)(b–1)	$\frac{SS_{A\times B}}{d.f._{A\times B}}$
błędu (*wewnątrzgrupowa*)	$SS_{\text{całk}} - SS_{\text{grup}}$	$\sum_{i=1}^{a}\sum_{j=1}^{b}(n_{ij}-1)$ lub $d.f._{\text{całk.}} - d.f._{\text{grup}}$	$\frac{SS_{błędu}}{d.f._{błędu}}$

a jest liczbą grup czynnika A, podobnie b jest liczbą grup czynnika B, zaś n_{ij} jest liczbą powtórzeń w komórce ij macierzy dla i-tej grupy czynnika A i j-tej grupy czynnika B; wartość C obliczamy ze wzoru:

$$C = \frac{\left(\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{l=1}^{n_{ij}} X_{ijl}\right)^2}{N} \qquad N = \sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij}\,.$$

Jeżeli w którejś grupie brakuje pojedynczych danych (< 10%), to można je wygenerować według równania:

$$\hat{X}_{ijl} = \frac{aA_i + bB_j - \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{l=1}^{n_{ij}} X_{ijl}}{N+1-a-b},$$

gdzie A_i jest sumą innych danych w grupie *i*-tej czynnika A i B_j jest sumą innych danych w grupie *j*-tej czynnika B; $\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{l=1}^{n_{ij}} X_{ijl}$ jest sumą wszystkich występujących danych, *N* jest całkowitą liczbą danych (uwzględniając także dane brakujące) w modelu.

Dwuczynnikowa analiza wariancji z pojedynczymi pomiarami w grupie (model z równą liczebnością grup bez powtórzeń)
(balanced design without replication)

Metoda ta, która porównuje wartości więcej niż dwóch zmiennych mierzonych u tego samego osobnika, jest rozwinięciem sparowanego testu *t* Studenta. Jest ona szczególnym przypadkiem metody analizy zrandomizowanych bloków. Różnice między tymi dwoma podejściami ilustrują przykłady w *„Części II – Uzupełnienia, przykłady i zadania"* niniejszego opracowania.

Sumy kwadratów dla tego modelu obliczamy w następujący sposób:

zmienność	**suma kwadratów (*SS*)**	**stopnie swobody (*d.f.*)**	**błąd średnio-kwadratowy (*MS*)**
całkowita	$\sum_{i=1}^{a}\sum_{j=1}^{b} X_{ij}^2 - C$	N–1	
czynnika 1 (A)	$\frac{\sum_{i=1}^{a}\left(\sum_{j=1}^{b} X_{ij}\right)^2}{b} - C$	a–1	$\frac{SS_A}{d.f._A}$
czynnika 2 (B)	$\frac{\sum_{j=1}^{b}\left(\sum_{i=1}^{a} X_{ij}\right)^2}{a} - C$	b–1	$\frac{SS_B}{d.f._B}$
resztowa	$SS_{\text{całk}} - SS_{\text{A}} - SS_{\text{B}}$	(a–1)(b–1)	$\frac{SS_{reszt}}{d.f._{reszt}}$

a jest liczbą grup czynnika A, *b* jest liczbą grup czynnika B; wartość C obliczamy z równania:

$$C = \frac{\left(\sum_{i=1}^{a}\sum_{j=1}^{b} X_{ij}\right)^2}{N}, \quad \text{gdzie } N = ab.$$

Zauważmy, że w metodzie tej nie mamy miary interakcji między czynnikami, ani też miary zmienności błędu. Zakładamy, że interakcja – jeżeli ją wykrywamy – może wynikać jedynie ze zmienności losowej (być dziełem przypadku), którą wyraża wariancja resztowa (*residual variability*). Możliwość występowania rzeczywistej interakcji między czynnikami ma wpływ na moc wnioskowania.

Jeżeli założymy, że mogą wystąpić interakcje między czynnikami to postępujemy tak:

badany efekt	**model 1 (oba czynniki: A i B stałe)**	**model 2 (oba czynniki: A i B losowe)**	**model 3 (czynnik A stały: czynnik B losowy)**
czynnik A	$\frac{MS_A}{MS_{A\times B}}$ *	$\frac{MS_A}{MS_{reszt}}$	$\frac{MS_A}{MS_{reszt}}$
czynnik B	$\frac{MS_A}{MS_{A\times B}}$ *	$\frac{MS_B}{MS_{reszt}}$	$\frac{MS_B}{MS_{błędu}}$ *
interakcja A x B	nie możemy badać	nie możemy badać	nie możemy badać

* $MS_{A\times B}$ liczone jak w przypadku dwuczynnikowej ANOVA z powtórzeniami; zwiększone ryzyko popełnienia błędu II rodzaju (β).

Jeżeli natomiast założymy, że nie ma rzeczywistej interakcji między czynnikami, to mamy do wyboru następujące możliwości:

badany efekt	**model 1 (oba czynniki: A i B stałe)**	**model 2 (oba czynniki: A i B losowe)**	**model 3 (czynnik A stały: czynnik B losowy)**
czynnik A	$\frac{MS_A}{MS_{reszt}}$	$\frac{MS_A}{MS_{reszt}}$	$\frac{MS_A}{MS_{reszt}}$
czynnik B	$\frac{MS_B}{MS_{reszt}}$	$\frac{MS_B}{MS_{reszt}}$	$\frac{MS_B}{MS_{reszt}}$
interakcja A x B	nie możemy badać	nie możemy badać	nie możemy badać

Jest to metoda, która ma bardzo często odzwierciedlenie w praktyce badawczej, kiedy jedną próbkę/osobnika testujemy przy użyciu różnych technik, lub poddajemy działaniu różnych czynników.

Jest to prostszy wariant modelu badawczego, gdzie mamy do czynienia ze sparowanymi pomiarami w wielu grupach. Kiedy na przykład porównujemy wartości badanego parametru w grupie kontrolnej oraz kilku grupach badanych poddawanych działaniu jakiegoś czynnika – każdy pomiar „*kontrolny*" jest wtedy w jakiś sposób skojarzony z każdym z pomiarów odzwierciedlających wynik działania badanego czynnika. W takich sytuacjach często wykorzystujemy tzw. model analizy zrandomizowanych bloków (*randomized block design*) (ANOVA 3 bez powtórzeń – *mixed model without replication*) (zobacz też w *„Części II – Uzupełnienia, przykłady i zadania"*).

Porównania wielokrotne

W przypadku badania różnic między więcej niż dwoma średnimi prosta analiza wariancji dostarcza nam odpowiedzi na pytanie, czy średnie w poszczególnych porównywanych grupach są identyczne, czy też różnią się. Kiedy odrzucamy hipotezę zerową i przyjmujemy hipotezę alternatywną mówiącą, iż średnie w badanych grupach nie są takie same, nasza ciekawość nie jest w pełni zaspokojona, gdyż nadal nie wiemy, które średnie odróżniają się od siebie, a które są identyczne. Zauważmy, że nawet różnica (choć odpowiednio duża) w zakresie jednej z kilku porównywanych grup może nas przywieść do wartości testu *F* istotnie większej od 1, co uprawni nas do odrzucenia hipotezy zerowej. W tym miejscu kończy się użyteczność metody analizy wariancji, gdyż w dalszym postępowaniu metoda ta w czystej postaci nie jest w stanie zweryfikować indywidualnych porównań średnich poszczególnych grup, np. grupy nazwanej kontrolną z każdą z wielu grup badanych. W takich sytuacjach korzystamy z testów porównań wielokrotnych lub poprawki Bonferroniego dla testu *t* Studenta. Jeszcze innych możliwości dostarcza nam wywodząca się z analizy wariancji analiza kontrastów, która umożliwia testowanie statystycznej istotności prognozowanych szczegółowych różnic w określonych fragmentach naszego bardziej złożonego modelu (np. przy badaniu wpływu polimorfizmu genetycznego oraz występowania jednostki chorobowej na oporność na lek może się okazać, że polimorfizm wpływa na oporność, ale tylko w przypadku występowania choroby).

Testowanie istotności różnic między więcej niż dwoma grupami przy użyciu testu t Studenta z poprawką Bonferroniego

Jednym z najczęściej popełnianych w praktyce błędów jest stosowanie prostego testu *t* Studenta do takich indywidualnych porównań. Na czym polega ten błąd, niech nam pokaże poniższy przykład. W klasycznym ujęciu nasza hipoteza zerowa precyzuje, że średnie w porównywanych grupach są takie same. Hipotezę tę weryfikujemy przy poziomie istotności α, czyli ryzykujemy z prawdopodobieństwem α, że odrzucimy hipotezę zerową wtedy, gdy jest ona *de facto* prawdziwa. Na przykład, kiedy porównujemy trzy grupy, nasza globalna hipoteza zerowa będzie miała postać: H_0: $\mu_1=\mu_2=\mu_3$, podczas gdy dla indywidualnych hipotez zerowych możemy rozpisać następujące warianty porównań:

$H_{01}: \mu_1 = \mu_2$
$H_{02}: \mu_1 = \mu_3$
$H_{03}: \mu_2 = \mu_3$.

Aby hipoteza globalna pozostawała prawdziwa, muszą być prawdziwe wszystkie hipotezy „cząstkowe". Jeżeli przyjęty poziom istotności wynosi $\alpha = 0.05$, oznacza to, że prawdopodobieństwo, iż pomylimy się odrzucając taką hipotezę wynosi 5%, a prawdopodobieństwo, że się nie pomylimy odpowiednio 95%. Aby był spełniony warunek hipotezy zerowej, musielibyśmy nie pomylić się testując H_{01}: $\mu_1 = \mu_2$ oraz nie pomylić się testując H_{02}: $\mu_1 = \mu_3$, a także nie pomylić się testując H_{03}: $\mu_2 = \mu_3$. Wypadkowe prawdopodobieństwo spełnienia tych trzech warunków wynosi: $0.95 \times 0.95 \times 0.95 = (0.95)^3 = 0.8574$ lub ogólniej $(1-\alpha)^n$, gdzie *n* oznacza liczbę porównań. Odpowiednio, prawdopodobieństwo, że pomylimy się odrzucając hipotezę globalną wynosiłoby $1-0.8574 = 0.1426$, czyli blisko 15% – o wiele więcej niż najwyższy dopuszczalny poziom istotności ($\alpha = 0.05$).

Kiedy testujemy $i = N$ niezależnych indywidualnych hipotez zerowych H_{0i}, każdą na poziomie α, to prawdopodobieństwo błędnego odrzucenia H_0 będzie wielokrotnością α. To znaczy, jeżeli istotność jednego lub wielu z N indywidualnych testów, każdego na poziomie istotności α, będzie oznaczała odrzucenie globalnej hipotezy zerowej, to ryzyko popełnienia błędu przy takim odrzuceniu w przypadku, gdy globalna H_0 jest prawdziwa, wynosi o wiele więcej niż α. Dla $N = 2$ statystyk porównań indywidualnych, przy założeniu, że jedna jest niezależna od drugiej, prawdopodobieństwo odrzucenia przynajmniej jednej z dwóch hipotez zerowych H_{0i}, $i = 1,2$, każdej testowanej przy poziomie istotności $\alpha = 0.05$, w przypadku gdy obie są naprawdę prawdziwe, jest takie samo, jak wyrzucenie przynajmniej jednej dwudziestki podczas rzucania dwoma 20-ściennymi kostkami. Prawdopodobieństwo (P) to wynosi: P(wyrzucenia 20-stki pierwszą kostką i wyrzucenia nie-20-stki drugą kostką) lub P(wyrzucenia nie-20-stki pierwszą kostką i wyrzucenia 20-stki drugą kostką) lub P(wyrzucenia 20-stki pierwszą kostką i wyrzucenia 20-stki drugą kostką):

$$\left(\frac{1}{20}\times\frac{19}{20}\right)+\left(\frac{19}{20}\times\frac{1}{20}\right)+\left(\frac{1}{20}\times\frac{1}{20}\right)=\frac{19}{400}+\frac{19}{400}+\frac{1}{400}=\frac{39}{400}=0.0975$$

czyli jest prawie dwukrotnie większe niż dla każdej z pojedynczych indywidualnych hipotez.

Gdyby obie indywidualne hipotezy były wzajemnie wykluczające się, to prawdopodobieństwo odrzucenia globalnej H_0 stosując tę procedurę byłoby podwojeniem wartości α ($2 \times 0.05 = 0.10$). Przeciwnie, gdyby były one całkowicie zależne, prawdopodobieństwo odrzucenia prawdziwej H_0 wynosiłoby $\alpha = 0.05$. W rzeczywistości przyjmuje ono wartości pośrednie między 0.05 i 0.1, gdyż w praktyce hipotezy takie zależą od siebie do pewnego stopnia, choć niecałkowicie.

Uogólniając, kiedy stosujemy poziom istotności α^* do testowania N hipotez indywidualnych składających się na hipotezę globalną, wtedy bezpieczna górna granica prawdopodobieństwa odrzucenia prawdziwej globalnej hipotezy zerowej powinna wynosić $\alpha = N\alpha^*$, co definiuje $\alpha^* = \alpha/N$ jako bezpieczny poziom istotności dla każdej z indywidualnych hipotez zerowych. Jest to tak zwana poprawka Bonferroniego dla obliczania poziomu istotności α w metodach opartych na statystyce testu t Studenta. Tabela 3 przedstawia przykładowe skorygowane wartości krytyczne $t_{0.05(2),\infty}$ dla przypadków, gdy liczba grup zmiennej dyskretnej jest większa od 2. Poprawka ta opiera się na założeniu, że do odrzucenia globalnej hipotezy zerowej wystarczy wykazać nieprawdziwość jedynie jednej z hipotez indywidualnych. Istnieją alternatywne podejścia do tego przedstawionego powyżej, ale z uwagi na ogólny charakter niniejszego opracowania nie będą one tutaj szerzej dyskutowane.

Tab. 3. Poprawka Bonferroniego dla poziomu istotności $\alpha = 0.05$ przy różnej liczbie grup zmiennej kategoryzującej (dyskretnej).

liczba grup zmiennej dyskretnej	**liczba możliwych sparowanych porównań**	**poprawka Bonferroniego**	**skorygowana wartość krytyczna** $t_{0.05(2),\infty}$
2	1	0.05	1.96
3	3	0.0167	2.39
4	6	0.0083	2.65
5	10	0.005	2.81
6	15	0.0033	2.96
7	21	0.0024	3.08
8	28	0.0018	3.15
9	36	0.0014	3.23
10	45	0.0011	3.30

Należy zauważyć, że uproszczenie polegające na stosowaniu poprawki Bonferroniego zamiast testów porównań wielokrotnych sprawdza się w przypadkach dużych liczebności prób, przy założeniu, że wartości statystyki *t* odczytywane są dla nieskończonej liczby stopni swobody (tak jak ma to miejsce dla wartości krytycznych *z* testu normalnego). Możemy zauważyć, jak bardzo wzrasta wartość krytyczna (*kolumna 4*) ze wzrostem liczby grup zmiennej dyskretnej (grupującej, *kolumna 1*) (jako odzwierciedlenie ryzyka błędu doświadczenia). Dla wysokich liczebności prób, wartości statystyki testu *t* są bardzo bliskie wartościom statystyki *z* testu normalnego (dla nieskończonej liczebności), np. $t_{0.05(2),100}$ wynosi 1.984 zaś $t_{0.05(2),\infty} = 1.96$.

Czym „*mniej nieskończona*" jest liczba obserwacji – tak jak to bywa w rzeczywistej praktyce badawczej – tym rzeczywista wartość α jest mniejsza od tej podanej w kolumnie 3 Tabeli.

Testy istotności porównań wielokrotnych
(significance tests for multiple comparisons)

Testy te stosuje się jako następstwo jednoczynnikowej analizy wariancji w celu bliższego scharakteryzowania wzajemnych różnic pomiędzy średnimi porównywanych grup. Duża rozmaitość testów porównań wielokrotnych może utrudniać wybór początkującemu badaczowi, ale najczęściej stosowane są dwie z tych metod – przede wszystkim z uwagi na ich dużą moc i stosunkową niewrażliwość na naruszenie założeń normalności i jednorodności wariancji – test Tukeya, oraz test Newmana-Keulsa. Niektórzy stosują także chętnie test Dunnetta oraz test najmniejszych istotnych różnic (test NIR, *least significant difference* (*LSD) test*). Pierwszy nadaje się zwłaszcza do porównywania szczególnych par średnich (np. kontroli i grupy badanej), drugi jest ceniony szczególnie wtedy, gdy liczba wszystkich grup jest duża, a wartości średnie nie są bardzo różne. Na uwagę zasługuje także test do analizy kontrastów wielokrotnych (test Scheffe'go, test S), stosowany dość rzadko z uwagi na mniejszą moc niż np. test Tukeya, ale przydatny do szczególnych zestawień porównań wielokrotnych.

W swoim założeniu testy porównań wielokrotnych wymagają normalności rozkładu oraz jednorodności wariancji – podobnie jak analiza wariancji – ale test Tukeya jest słabo wrażliwy na naruszenie tych wymagań. Tak jak we wszystkich przypadkach, szczególnie kłopotliwe w przypadku większości testów porównań wielokrotnych wydaje się naruszenie

tego drugiego założenia. Ta niewrażliwość – jak i zresztą moc tych testów – zależy bardzo silnie od liczebności grup, jak i od tego, czy grupy są jednakowo liczne.

Test Tukeya jest testem o największej mocy wśród wszystkich testów porównań wielokrotnych, to znaczy ryzyko nieodrzucenia hipotezy zerowej, w przypadku kiedy jest ona naprawdę fałszywa, jest najmniejsze. W obu testach (Tukeya i Newmana-Keulsa) logika obliczeń jest bardzo podobna i opiera się na oszacowaniu rozstępu między średnimi, a następnie testowaniu różnicy rozstępów dla danej liczebności grupy. Różnica między obu testami z rachunkowego punktu widzenia polega na tym, że w teście Newmana-Keulsa przed testowaniem różnicy rozstępów, najpierw sortuje się średnie w porządku rosnącym.

Przykłady wykorzystania testu Tukeya w wariantach z równą i różną liczebnością grup, testu Newmana-Keulsa, testu Dunnetta oraz testu Scheffe'go może Czytelnik znaleźć w *„Części II – Uzupełnienia, przykłady i zadania"*.

Istnieją także nieparametryczne warianty tych testów, które można zastosować w sytuacjach trudności w ominięciu poważnych naruszeń założeń stosowania metod parametrycznych.

Testy badania jednorodności wariancji

Testy istotności do porównania dwóch lub więcej grup opierają się na założeniu, że wariancje w różnych grupach są takie same (tzn., że grupy te są jednorodne). O populacjach, których wariancje nie różnią się istotnie między sobą mówimy, że są ***homoscedastyczne***, zaś takie, których wariancje są różne nazywamy ***heteroscedastycznymi.*** Do badania równości wariancji służą testy jednorodności wariancji; do najczęściej stosowanych należą testy Fishera-Snedecora, Bartletta, czy Levene'a. Ten ostatni test, który jest szczególnie chętnie stosowany, jest równoważny z przeprowadzeniem jednokierunkowej ANOVA na wynikach odchyleń standardowych (od średnich w porównywanych grupach). Jego logika opiera się na tym, że im większa wariancja w obrębie grupy, tym większe są bezwzględne odchylenia od odpowiedniej średniej. Jeżeli statystyka *F* testu Levene'a okaże się istotna, to hipotezę o jednorodności wariancji należy odrzucić.

Przykłady zastosowania niektórych z nich znajdzie Czytelnik w *„Części II – Uzupełnienia, przykłady i zadania"*.

Podsumowanie

- W praktyce możemy napotkać sytuacje, kiedy liczba istotnych dla nas czynników będzie wyższa niż dwa. Z drugiej strony, nasz model może uwzględniać kilkakrotne pomiary wartości jakiejś cechy. Dokładny opis różnych wariantów analizy statystycznej takich szczególnych przypadków znajdzie Czytelnik w znakomitych opracowaniach statystycznych podanych w wykazie piśmiennictwa.

O czym powinniśmy pamiętać przed przystąpieniem do opracowania danych metodą analizy wariancji?

- Obejrzeć jak wygląda rozkład wyników, można przeprowadzić odpowiednie testy normalności czy jednorodności wariancji, pamiętając jednak, że spełnienie założeń normalności i homoscedastyczności nie jest bezwzględnie krytyczne dla zastosowania ANOVA.
- Aby nasze dane lepiej przystawały do tych wymagań, możemy zastosować transformację danych pomiarowych.
- Bardzo pożądane jest właściwe zaplanowanie doświadczenia, szczególnie, gdy badamy wpływ więcej niż jednego czynnika (zmiennej grupującej/kategoryzującej), tak abyśmy już przed rozpoczęciem doświadczenia mieli świadomość, z jakiego modelu analizy skorzystamy przy opracowywaniu wyników.
- Jeżeli stwierdzimy występowanie istotnej interakcji między czynnikami, to bezzasadne staje się określanie wpływu każdego z czynników, gdyż ujawnionego efektu i tak nie będziemy w stanie przypisać konkretnemu czynnikowi; możemy jednak wtedy zbadać łączny wpływ obu czynników badając zmienność grupową.

Testy istotności do oceny biozgodności (leków)

Omówione dotychczas metody znalazły zastosowanie w naukach farmakologicznych i farmaceutycznych do badania biodostępności oraz biozgodności leków. Biodostępnością (*bioavailability*) określamy szybkość oraz ilość leku, który staje się dostępny w aktywnej formie w miejscu jego działania w ustroju. Biozgodnością (*bioequivalence*) natomiast nazywamy wskazanie, że różne leki, preparaty lub różne stężenia tego samego preparatu stają się dostępne w tym samym czasie w tej samej ilości. Deklaracja biozgodności, czyli równej biodostępności oznacza, że leki lub preparaty mogą być stosowane wymiennie, gdyż prowadzą do tego samego skutku terapeutycznego.

Testy do weryfikowania biozgodności (*bioequivalence tests*) tym się różnią od dotychczas omawianych testów statystycznych, że ich celem jest udowodnienie braku różnic, nie zaś wykazanie występowania takich różnic. Ponieważ istnieje ścisła zależność między stężeniem leku we krwi a jego stężeniem w docelowym miejscu działania w ustroju, w procesach określania biozgodności opieramy się na określaniu wartości trzech podstawowych parametrów: maksymalnego stężenia leku w osoczu krwi (C_{max}), czasu osiągania stężenia maksymalnego w osoczu (T_{max}) oraz tzw. powierzchni pod krzywą absorpcji leku (AUC, *area under curve*) w docelowych tkankach w ustroju.

Do badania biozgodności stosuje się różne modele badawcze; do najbardziej popularnych należy badanie różnych formuł w niezależnych grupach ochotników o wyrównanej liczebności lub tzw. model krzyżowy. W tym pierwszym schemacie badawczym ochotników przydziela się w sposób losowy do jednej z porównywanych grup. Ponieważ model ten zakłada stosunkowo niewielką zmienność międzyosobniczą w odpowiedzi na dany lek, jest on rzadziej stosowany. W modelu krzyżowym ochotników także przydziela się pierwotnie w sposób losowy do jednej z grup badanych, ale każda z osób włączonych do badań otrzymuje każdy z badanych leków w ciągu trwania całego badania. Oczywiście taki schemat postępowania zakłada istnienie okresów niepodawania żadnego z leków (*washout period*) w celu wykluczenia możliwych efektów interferencji kilku preparatów. Jest sprawą umowną, jak długa powinna być taka przerwa: zależy to od rodzaju preparatu/leku,

standardowo przyjmuje się, że powinien on być nie krótszy niż pięciokrotny okres czasu połowicznej degradacji preparatu w ustroju. Ponieważ często nie znamy wartości takiego okresu połowicznej degradacji (szczególnie w przypadku nowo badanego preparatu), możemy oszacować długość okresu eliminacji leku z ustroju jedynie z pewnym przybliżeniem. Taki schemat krzyżowy badania biozgodności góruje nad pierwszym schematem w niezależnych grupach badanych przede wszystkim tym, że uwzględnia występowanie dużej zmienności międzyosobniczej; jest to wymierna korzyść w większości badań klinicznych. Pozwala on ponadto w przybliżeniu dwukrotnie ograniczyć liczbę ochotników włączonych do badania, gdyż każda osoba badana stanowi jednocześnie swoją własną kontrolę, a zatem ogranicza się w ten sposób zmienność wewnątrzosobniczą w próbie, co wpływa z kolei na utrzymanie mocy testu przy niższej liczebności próby badanej.

Ocena biozgodności polega zasadniczo na niewykazaniu różnic między biodostępnością dwóch (lub kilku) różnych preparatów. Ponieważ najpowszechniej i najchętniej stosowanymi testami przy badaniu różnic są testy oparte na statystyce testu *t* Studenta, również i w tym wypadku wykorzystuje się logikę tych procedur, ale z podkreśleniem, że nasze podejście do testowania słuszności hipotez w badaniu biozgodności jest diametralnie różne niż przy testowaniu istotności różnic w klasycznym zastosowaniu testu *t*. Jest tak dlatego, że klasyczne testy istotności są nastawione na wykazanie różnic między wynikami, nie zaś na wykazanie równości. Tej równości, jak pamiętamy, nie jesteśmy w stanie nigdy dowieść, nawet, jeżeli nie udaje nam się odrzucić hipotezy zakładającej występowanie takiej równości wyników i przyjąć hipotezy zakładającej istnienie różnicy. W testach oceny biozgodności nasz cel jest diametralnie inny: pragniemy dowieść istnienia równości i odrzucić hipotezę zakładającą występowanie różnic. Pamiętamy również, że wszystkie testy parametryczne, w tym, także testy *t*, zakładają normalny, a więc m.in. symetryczny, rozkład danych doświadczalnych. Jednakże większość danych dotyczących biodostępności charakteryzuje się rozkładem odbiegającym od normalnego i wykazującym prawoskośność. W takich sytuacjach musimy uciekać się do transformacji danych, a skoro rozkład wykazuje prawoskośność, to dobrym wyborem jest transformacja logarytmiczna.

Kiedy testujemy próby o małej liczebności, nasza szansa nieodrzucenia hipotezy zerowej zakładającej, że próby nie różnią się jest większa, szczególnie, jeżeli zmienność międzyosobnicza jest duża. Zatem badając mało liczne próby często mielibyśmy szansę niewykazania różnic, chociaż to nie to samo, co wykazanie równości. W sytuacji, kiedy pragniemy wykazać brak różnic – tak jak ma to miejsce w testach biozgodności – mogłoby to się wydawać kuszące, gdyby nie fakt, że przy małej liczebności prób maleje także moc testu. Dlatego też, aby zobiektywizować wyniki testów biozgodności ustalamy pewien minimalny poziom mocy wnioskowania oraz minimalny zakres różnicy między porównywanymi preparatami, który uznalibyśmy za istotny statystycznie. W praktyce, przy orzekaniu biozgodności między badanymi preparatami/lekami przyjmuje się zwykle, że próba musi być wystarczająco liczna, aby z mocą przynajmniej 80% orzekać o istotności 20% różnicy. Oznacza to, że biodostępność testowanego preparatu musi się mieścić w zakresie od 80% do 120% biodostępności dla preparatu referencyjnego. Tą 20% różnicę możemy testować stosując obliczanie przedziałów ufności dla testów jedno- lub obustronnych:

$$\mu_{testowana} - \mu_{referencyjna} = (\bar{x}_{testowana} - \bar{x}_{referencyjna}) \pm t\sqrt{\frac{2s_p^2}{n}},$$

$$\text{czyli} \quad \mu_{testowana} - \mu_{referencyjna} = d \pm t^* SE$$

gdzie $d = \bar{x}_{testowana} - \bar{x}_{referencyjna}$ oznacza różnicę między preparatami,

zaś $s_p^2 = \frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2)}$ oraz $\sqrt{\frac{2s_p^2}{n}} = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$.

Dla przyjmowanego zakresu 80%-120% biodostępności możemy zapisać:

$$0.80 < \mu_{testowana} - \mu_{referencyjna} < 1.20 \quad \text{lub} \quad 80\% < \frac{\mu_{testowane}}{\mu_{referencyjne}} < 120\% \ ,$$

gdzie 80% i 120% stanowią dolny i górny kres przyjmowanego przedziału ufności. Wartości kresu dolnego i kresu górnego możemy obliczyć następująco:

$$\text{kres dolny} = \frac{(d - SE) + \bar{x}_{referencyjne}}{\bar{x}_{referencyjne}} \times 100\%$$

$$\text{kres górny} = \frac{(d + SE) + \bar{x}_{referencyjne}}{\bar{x}_{referencyjne}} \times 100\%$$

Ostatecznie nasz przedział ufności będzie wynosił:

$$\text{kres dolny} < \frac{\mu_{testowane}}{\mu_{referencyjne}} < \text{kres górny.}$$

Nie powinno dziwić, że o biodostępności decyduje istotnie profil rozpuszczalności preparatu, ponieważ lepsza rozpuszczalność w środowisku płynów ustrojowych warunkuje lepszą penetrację i podaż leku do wybranych tkanek w organizmie.

Wyrażając stopień biozgodności lub „*bio-niezgodności*" pod względem takiego profilu rozpuszczalności porównywanych preparatów, podaje się często wartości tzw. współczynników różnicy ($f_{różnicy}$) oraz współczynników podobieństwa ($f_{podobieństwa}$). Pod względem rachunkowym współczynniki te można wyrazić za pomocą następujących równań:

$$f_{różnic} = \frac{\sum |R_t - T_t|}{\sum R_t} \times 100\%$$

$$f_{podobieństw} = 50 \log \left[\frac{1}{\sqrt{1 + \frac{1}{n} \sum (R_t - T_t)^2}} \times 100\% \right],$$

gdzie n oznacza liczbę testowanych punktów (czasowych) profilu rozpuszczalności, R_t i T_t oznaczają procentowe frakcje rozpuszczone po czasie t substancji referencyjnej i testowanej.

Oczekujemy, że dla preparatów wysoce biozgodnych wartość $f_{różnicy}$ powinna być bliska zeru (a przynajmniej niższa niż 15%), zaś wartość $f_{podobieństwa}$ bliska 100% (a przynajmniej wyższa niż 50%). Dla dwóch identycznych lub niemal identycznych profili rozpuszczalności wartość $f_{podobieństwa}$ wynosi około 100%. Czym niższa jest ta wartość, tym większa różnica występuje między dwoma porównywanymi profilami rozpuszczalności. Dwa podstawowe założenia tego testu, to: współczynnik zmienności nieprzekraczający 10-20% oraz niewłączanie danych o wartościach średnich zmiennej zależnej istotnie powyżej 85% wartości maksymalnej.

Chociaż opisane tutaj procedury opracowano z myślą o ocenie biozgodności leków, to mogą być one z powodzeniem stosowane do szacowania zgodności między różnymi grupami (poziomami) jakichkolwiek zmiennych o rozkładach dyskretnych. Współczynniki podobieństwa i różnic zostały stworzone jako miary do oceny biozgodności w profilach rozpuszczalności; zauważmy jednak, że mogą być one z powodzeniem wykorzystywane do oceny pokrywania się jakichkolwiek krzywych obrazujących zależności czasowe, stężeniowe, itp., bez potrzeby estymacji współczynników równań regresji nieliniowej opisujących te procesy.

Ocena zgodności dwóch metod (pomiarowych)

W praktyce badawczej spotykamy się często z problemem oceny zgodności dwóch różnych metod stosowanych w praktyce klinicznej, diagnostycznej, laboratoryjnej, itp. Nasze zainteresowanie zwracamy wtedy najczęściej w kierunku określenia, czy nowa metoda pomiaru (opracowana lub zaadoptowana przez nas jako alternatywna do już istniejącej i powszechnie stosowanej procedury) dostarcza wyników zgodnych z metodą wcześniej stosowaną i w związku z tym czy istnieje możliwość:

- wymiennego stosowania obu metod,
- zastąpienia starszej metody nową (lub np. droższej tańszą).

Praktyka uczy, że w przypadkach takich uciekamy się często do analizy korelacji, przyjmując, iż współczynnik korelacji jest miarą dobrej lub złej korespondencji wyników uzyskanych jedną metodą oraz tych uzyskanych za pomocą drugiej metody. Zauważmy jednak, że przy porównywaniu wyników uzyskanych dwoma różnymi metodami, tak naprawdę nie znamy wartości rzeczywistej mierzonego parametru. Możemy sobie wyobrazić, że zarówno jedna, jak i druga metoda może nieco przekłamywać te rzeczywiste wartości. W tym przypadku nie dokonujemy przecież kalibracji w oparciu o znane wartości (np. na podstawie dokonanych przez nas naważek, sporządzonych roztworów o określonym stężeniu, itp.) lub też wartości oznaczone za pomocą bardzo precyzyjnej referencyjnej aparatury, lecz porównujemy wyniki zebrane przy użyciu dwóch procedur (metod), o których dokładności czy precyzji nie mamy szczegółowych informacji, a opieramy się jedynie na danych mniej lub bardziej szacunkowych. W takim przypadku zależy nam na orzeczeniu, że obie metody dostarczają zgodnych wyników, nie zaś na udowodnieniu, iż są one dokładne (to znaczy, że generują wyniki zgodne z wartościami rzeczywistymi). Współczynnik korelacji odzwierciedla asocjację między parami wyników, a nie zgodność wyników każdej pary. Zauważmy, że korelacja pozostanie wysoka zawsze wtedy, gdy punkty o współrzędnych wyznaczających wartości w każdej z par (w naszym przypadku parę wyników tworzą wartości uzyskane jedną metodą lub drugą metodą) będą leżały na prostej, podczas gdy warunkiem wysokiej zgodności jest położenie punktów na przekątnej

układu współrzędnych (nachylonej pod kątem 45° w stosunku do osi odciętych). Z oczywistych względów, jeżeli skale pomiarowe obu porównywanych metod będą się różnić, możemy nadal rejestrować silną korelację, ale na pewno nie uzyskamy zgodności metod. Wydaje się zupełnie oczywiste, że skoro porównujemy dwie metody, które zaprojektowano do pomiaru tego samego parametru, oczekujemy wysokiej zależności między wynikami dostarczonymi przez te metody, a zatem wysokich wartości współczynnika korelacji, i tym samym wysokiej istotności statystycznej takiej korelacji. Nie oznacza to jednak, że wyniki zebrane przy użyciu jednej i drugiej metody są zgodne i że metody te można stosować wymiennie.

We wczesnych latach 80-tych Altman i Bland* zaproponowali prostą do interpretacji graficzno-rachunkową metodę oceny zgodności wyników zbieranych różnymi metodami. Zaletą tej metody jest, że możemy ją stosować zarówno do porównywania obserwacji zebranych dwoma metodami przez jednego badacza, wyników zebranych jedną metodą przez różnych badaczy, jak również do oszacowana powtarzalności pomiarów.

Oceniając stopień zgodności, chcemy się zwykle dowiedzieć, jak bardzo dwie metody różnią się od siebie. Oczywiście, należy się pogodzić z faktem, że wyniki uzyskane obiema metodami nigdy nie będą takie same, podobnie jak wyniki w serii zebrane tą samą metodą nie będą nigdy zupełnie identyczne. Jeżeli taka wykryta niezgodność nie jest duża, to możemy wnioskować, że jedna metoda może być zastąpiona przez drugą lub że obie można stosować wymiennie. Możemy tak postąpić, przyjmując, że te niewielkie obserwowane różnice nie będą miały istotnego wpływu na decyzje kliniczne, terapeutyczne, itp. dotyczące pacjenta, u którego przeprowadzano badanie/oznaczenie. Miary takiej zgodności/niezgodności są dość arbitralne i optymalnie powinny być dobrane *a priori* w celu ustalenia metody porównania oraz oszacowania liczebności próby.

Tym co nas interesuje przy rozważaniu zgodności porównywanych metod jest także powtarzalność każdej z nich. Jest tak dlatego, że stopień niepowtarzalności każdej z metod jest w oczywisty i naturalny sposób ograniczeniem zgodności tych metod. Jeżeli metoda dostarcza mało powtarzalnych obserwacji, czyli jeżeli zmienność wewnątrzgrupowa serii pomiarów przeprowadzonej dla tego samego obiektu badań jest duża, to oczywiście nie będziemy także oczekiwali wysokiej zgodności wyników tej metody z jakąkolwiek inną porównywaną metodą. Z tych samych powodów – jeżeli starsza metoda, którą chętnie zastąpilibyśmy nową doskonalszą metodą, jest bardzo mało precyzyjna, to naturalnie jej zgodność z nową metodą (nawet super precyzyjną) będzie znikoma.

Jeżeli zakładamy, że pomiary dotyczące tego samego obiektu badań są powtarzalne, to oczywiście oczekujemy, że średnia różnica między pomiarami będzie bliska zeru. Jeżeli założenie to nie jest spełnione, tzn. różnica jest istotnie różna od zera, wówczas nie jesteśmy w stanie wypowiedzieć się obiektywnie na temat powtarzalności pomiarów. Źródłem takiej niespójności może być to, że proces przeprowadzania pierwszego pomiaru wpływa na wartości drugiego pomiaru: na przykład znajomość wartości pierwszej obserwacji wpływa na rejestrowany wynik drugiej, lub przeprowadzenie pierwszego pomiaru zmienia istotnie obiekt badany (pomijamy przypadki, gdy wykorzystywana przez nas metoda charaktery-

* Obszerniejsze omówienie tej metody znajdzie Czytelnik w pracach: Altman D.G., Bland J.M.: Measurement in medicine: the analysis of method comparison studies. Statistician 1983, 32, 307-317; Bland J.M., Altman D.G.: Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986, 1, 307-310.

zuje się z zasady niezadowalającą powtarzalnością). W praktyce badawczej różnice dla poszczególnych par pomiarów nie będą oczywiście nigdy dokładnie równe zeru, będą się charakteryzowały określona zmiennością. Oczekujemy, że przynajmniej w 95% wszystkich przypadków takie obliczone różnice będą mniejsze niż wartość podwójnego odchylenia standardowego dla średniej wszystkich zanotowanych różnic. Miarą powtarzalności będzie wtedy tzw. współczynnik powtarzalności (*repeatability coefficient*)** równy:

$$s_{rep} = 2s_d = 2\sqrt{\frac{\sum_{i=1}^{n} d^2}{n}}$$

s_{rep} oznacza współczynnik powtarzalności,

s_{d} jest odchyleniem standardowym różnic dla par pomiarów,

$\sum_{i=1}^{n} d^2$ oznacza sumę kwadratów różnic dla par pomiarów,

zaś n jest liczbą wszystkich analizowanych różnic par pomiarów.

W przypadku włączenia do analizy większej liczby powtórzeń zamiast statystyki testu *t* Studenta stosujemy analizę wariancji (zobacz też Armitage, 1994).

Bardziej szczegółowe omówienie procedury oceny zgodności dwóch metod pomiarowych oraz oceny powtarzalności pomiarów znajdzie Czytelnik w *„Części II – Uzupełnienia, przykłady, zadania"*.

** Wg: British Standards Institution. Precision of test methods (BS 5497, part 1). London: BSI, 1979.

Rozdział 6

Testy istotności dla proporcji

Test istotności dla pojedynczej proporcji przy zastosowaniu statystyki rozkładu dwumianowego

W najprostszym przypadku do badania istotności pojedynczej proporcji możemy zastosować statystykę rozkładu dwumianowego (zobacz str. 48):

$$P(r\ A) = \frac{n!}{r!(n-r)!}\pi^r(1-\pi)^{n-r}$$

oraz obliczyć prawdopodobieństwo, że wynik znajdzie się w obszarze krytycznym pod krzywą rozkładu. Podobnie, jak to ma miejsce w przypadku analizy rozkładu normalnego lub rozkładu t Studenta, poziom istotności różnic będziemy rozumieli jako prawdopodobieństwo uzyskania wyniku przynajmniej tak skrajnego (to znaczy leżącego w obszarze krytycznym rozkładu dwumianowego), jak wynik zaobserwowany. I tutaj, tak jak to miało miejsce w powyższych rozkładach wykorzystywanych w testach parametrycznych, możemy mieć do czynienia z jednostronnym lub obustronnym obszarem krytycznym (zobacz *„Część II – Uzupełnienia, przykłady i zadania"*).

W przypadku rozkładów symetrycznych względem miary centralnej wartość $p = 0.5$, czyli teoretyczne prawdopodobieństwo tego, że u losowo wybranego osobnika w populacji stwierdzimy występowanie cechy A wynosi 50%.

Z tej samej metody szacowania prawdopodobieństw na podstawie powyższego równania możemy także skorzystać, gdy $p \neq 0.5$, czyli gdy rozkład dwumianowy jest asymetryczny (zobacz Ryc. 4). Zauważmy jednak, że w takim szczególnym przypadku prawdopodobieństwa dla przedziałów krytycznych po lewej i prawej stronie rozkładu nie są równe. Aby w takiej sytuacji obliczyć istotność dla testu obustronnego możemy postąpić dwojako. Możemy zsumować (*i*) prawdopodobieństwo dla obserwowanej wartości cechy i prawdopodobieństwa dla wartości bardziej skrajnych od tej obserwowanej po tej samej stronie rozkładu, oraz (*ii*) wszystkie te prawdopodobieństwa po przeciwnej stronie rozkładu, które są mniejsze od prawdopodobieństwa dla wartości obserwowanej. Alternatywnie, możemy podwoić wartość sumy prawdopodobieństw dla wartości przynajmniej tak skrajnych (lub bardziej) jak wartość obserwowana po tej samej stronie rozkładu. Wyniki uzyskane tymi dwoma metodami będą oczywiście różne, ale praktyka pokazuje, że nie ma to zasadniczego

wpływu na podejmowanie decyzji o odrzucaniu hipotezy zerowej. Żadna z tych metod nie ma przewagi nad drugą, ale to drugie podejście jest prostsze pod względem rachunkowym.

Aproksymacja normalna rozkładu dwumianowego

W przypadku prób o dużych liczebnościach, n, obliczanie prawdopodobieństw dla poszczególnych wartości obserwowanych (i tych bardziej skrajnych od nich) w obszarach krytycznych rozkładu dwumianowego przy zastosowaniu powyższego równania może okazać się żmudne i pracochłonne. W takich przypadkach w celu oszacowania istotności korzystamy z aproksymacji rozkładu dwumianowego do rozkładu normalnego. Gdy n wzrasta, rozkład dwumianowy staje się bardzo zbliżony do rozkładu normalnego (Ryc. 4). Przyjmuje się, że rozkład normalny jest wystarczająco dobrym przybliżeniem do rozkładu dwumianowego, jeżeli zarówno $n\pi$, jak i $n - n\pi$ są równe lub większe od 5. Aproksymowany rozkład normalny ma taką samą średnią i odchylenie standardowe jak rozkład dwumianowy (Tab. 2).

Testy istotności i przedział ufności przy zastosowaniu aproksymacji do rozkładu normalnego

Aproksymowany do rozkładu dwumianowego rozkład normalny może być stosowany w testach istotności dla proporcji tak samo, jak jest wykorzystywany w testach istotności dla średnich.

Test istotności dla pojedynczej proporcji

W teście tym wartość statystyki z wyliczana jest według równania:

$$z = \frac{p - \pi}{\sqrt{\{\pi(1-\pi)/n\}}}$$

gdzie p oznacza wartość proporcji, π – wartość prawdopodobieństwa wystąpienia cechy w populacji ogólnej i n – liczebność próby.

W celu lepszego przybliżenia charakterystyki rozkładu dwumianowego, który jest rozkładem dyskretnym, do rozkładu normalnego, który jest rozkładem ciągłym, należy stosować poprawkę na ciągłość:

$$z = \frac{|p - \pi| - 1/(2n)}{\sqrt{\{\pi(1-\pi)/n\}}}$$

Przedział ufności dla pojedynczej proporcji

Obliczanie dokładnych wartości przedziału ufności rozkładu dwumianowego dla różnych proporcji w próbach o różnych liczebnościach jest bardzo złożoną procedurą,

głównie z uwagi na asymetrię rozkładu dla $p \neq 0.5$. O wiele wygodniej jest stosować właściwość przybliżenia rozkładu dwumianowego do rozkładu normalnego (sprawdzającą się tym lepiej, im większa jest liczebność próby):

$$CI = p \pm z' \text{ x } SE, \quad SE = \sqrt{\{p(1-p)/n\}}$$

Zauważmy, że przybliżenie to nie zawiera poprawki na ciągłość; jest ono akceptowalne, jeżeli wartości zarówno *np*, jak i *n – np* wynoszą przynajmniej 10.

Test istotności dla porównywania dwóch proporcji

Test służący do porównywania dwóch proporcji:

$$p_1 = r_1/n_1, \quad p_2 = r_2/n_2,$$

ma ogólną postać:

$$z = \frac{p_1 - p_2}{SE_{(p_1 - p_2)}}$$

gdzie wartość błędu standardowego różnicy $p_1 - p_2$, przy hipotezie zerowej zakładającej, że porównywane proporcje są sobie równe i nie różne od proporcji w populacji ogólnej (czyli $\pi_1 = \pi_2 = \pi$), wynosi:

$$SE_{(p_1 - p_2)} = \sqrt{\{\pi(1-\pi)\,(1/n_1 + 1/n_2)\}}.$$

Proporcja w populacji ogólnej jest szacowana jako całkowita uśredniona proporcja dla obu porównywanych grup, czyli:

$$\pi \cong p = \frac{r_1 + r_2}{n_1 + n_2}.$$

Podobnie jak w przypadku testu istotności dla pojedynczej proporcji, aproksymacja do rozkładu normalnego jest lepsza po zastosowaniu poprawki na ciągłość. Równanie na wartość statystyki *z* przyjmuje wtedy postać:

$$z = \frac{|p_1 - p_2| - \{1/(2n_1) + 1(2n_2)\}}{\sqrt{\{p(1-p)(1/n_1 + 1/n_2)\}}}$$

$$\text{dla} \quad p = \frac{r_1 + r_2}{n_1 + n_2}$$

Test normalny jest dobrą aproksymacją przy ocenie istotności różnic proporcji, jeżeli:

$$n_1 + n_2 > 40, \text{ albo } n_1 p \geq 5, \quad n_1 - n_1 p \geq 5, \quad n_2 p \geq 5 \text{ oraz } n_2 - n_2 p \geq 5.$$

Jeżeli warunek ten nie jest spełniony, powinniśmy skorzystać z dokładnego testu Fishera (zobacz *„Tabele liczebności i statystyki oparte na charakterystyce testu chi²"*).

Przedział ufności dla różnicy między dwoma proporcjami

Obliczanie przedziału ufności dla różnicy między dwoma proporcjami w aproksymacji do rozkładu normalnego według równania:

$$CI = (p_1 - p_2) \pm (z' \times SE) \qquad SE = \sqrt{\{p_1 (1 - p_1)/n_1 + p_2(1 - p_2)/n_2\}}$$

może być z powodzeniem stosowane, jeżeli tylko $n_1 p \geq 10$, $n_1 - n_1 p \geq 10$, $n_2 p \geq 10$ oraz $n_2 - n_2 p \geq 10$; tak jak poprzednio, aproksymacja taka jest tym lepsza, im wyrażenia te są wyższe od 10.

Rozdział 7

Metody estymacji liczebności próby

Zasadniczym wymaganiem w planowaniu badań naukowych jest oszacowanie wielkości próby, jaką zamierzamy przebadać. Robimy to m.in. po to, aby nie zbierać niepotrzebnie dużej liczby danych w sytuacji, gdy na przykład:

- dostrzegamy już na „pierwszy rzut oka", że porównywane grupy różnią się między sobą,
- nie występują rzeczywiste różnice i nie wykażemy ich niezależnie od liczebności próby.

Panuje powszechne przekonanie, że estymacja liczebności próby jest dobrym obyczajem w nauce oraz że stosowanie tej procedury powinno być nawykiem każdego rzetelnego badacza.

Niestosowanie metody estymacji wielkości próby jest jednym z czynników, które postrzegane są wręcz jako niekompetencja w prowadzeniu badań naukowych. Niestety doświadczenie uczy, że ocena liczebności grupy badanej przed wykonaniem badań jest bardzo rzadko stosowaną praktyką, a liczebność taka oceniana jest na czysto arbitralnych zasadach. Jest to praktyka uważana przez licznych badaczy za nieetyczną. Wykonując niepotrzebnie bardzo dużą liczbę powtórzeń nie tylko mnożymy niepotrzebnie koszty eksperymentu, podczas gdy moglibyśmy wykorzystać te środki na sprawdzenie innej koncepcji badawczej. W badaniach klinicznych wiąże się to nie tylko z podawaniem większej liczbie osób *placebo*, ale także realne opóźnianie wprowadzania do praktyki klinicznej korzystnej strategii farmakologicznej.

U niedoświadczonego badacza stosowanie procedur szacowania wielkości badanej próby zmierzających do maksymalnego ograniczenia jej niezbędnej liczebności może zrodzić wątpliwość: dlaczego nie możemy przebadać wystarczająco licznej próby, aby mieć olbrzymią pewność, że próba ta jest bardzo dobrze reprezentatywna dla badanej populacji. Absurdalność takiego pomysłu może wyjaśnić następujący przykład. Przypuśćmy, że chcemy zbadać twardość tabletek. Metoda badania twardości polega na rozłupywaniu tabletek przy użyciu specjalnego przyrządu i badania siły nacisku urządzenia. Mamy do dyspozycji 50000 tabletek. Jeżeli rozłupiemy wszystkie 50000 tabletek to oceniana przez nas średnia będzie równa średniej populacji generalnej – wiarygodność takiej estymacji będzie absolutna. Czy są jednak racjonalne przesłanki do badania wszystkich 50000 tabletek? Przecież jeżeli to zrobimy, to nie pozostanie nam już żadna tabletka do przeprowadzenia jakichkolwiek innych badań.

Metody estymacji liczebności próby opierają się na kilku założeniach:

1. Próby posiadają rozkład normalny – gdy liczebność próby bardzo wzrasta, wówczas średnie prób podlegają rozkładowi normalnemu nawet w sytuacji, gdy odpowiednia

zmienna w populacji nie posiada rozkładu normalnego lub nie jest wystarczająco dobrze zmierzona.
2. Musimy zdefiniować, z jakim prawdopodobieństwem pragniemy orzec o występowaniu lub braku różnic.
3. Estymowana liczebność zależy od mocy stosowanego testu, czyli musimy założyć jak duże ryzyko błędu II rodzaju (prawdopodobieństwo nieodrzucenia hipotezy zerowej, gdy jest ona fałszywa) dopuszczamy.

Niekiedy oszacowanie właściwej wielkości próby badanej wymaga bardzo precyzyjnego określenia, czego spodziewamy się (jakiej informacji) po wynikach badania. Na przykład w badaniach porównania śmiertelności wśród niemowląt karmionych odżywką w stosunku do tych karmionych piersią samo stwierdzenie większego ryzyka nie zadowala nas – pragniemy jeszcze wiedzieć, ile razy ryzyko takie jest większe. Wielkość próby będzie na przykład inna w przypadku 4-krotnego i dwukrotnego ryzyka. Należy także pamiętać, że z uwagi na zmienność wyników (widoczną szczególnie wyraźnie w przypadku małych prób) obserwowany wzrost ryzyka może być za mały, aby wykazać jego istotność. Dlatego powinniśmy *a priori* określić prawdopodobieństwo, z jakim chcielibyśmy wnioskować o istotności różnic na danym poziomie istotności, czyli powinniśmy ustalić moc wnioskowania.

W ten sposób możemy na przykład określić, że badanie dostarcza wartościowych wyników, jeżeli z prawdopodobieństwem 90% możemy stwierdzić, że ryzyko względne śmierci niemowląt karmionych butelką w stosunku do tych karmionych piersią jest na przyjętym poziomie istotności (np. 5%) przynajmniej tak wysokie jak 2.

W Tabeli 4 podane są wzory obliczania minimalnej liczebności próby dla różnych przypadków. W pierwszej części tabeli zamieszczono przypadki liczenia istotności różnic między grupami. W drugiej części podano sposoby szacowania określonej wartości z żądaną precyzją. Dla porównywania średnich, proporcji lub częstości szacowana liczebność odnosi się do każdej z grup; liczebność tę należy podwoić dla modeli doświadczalnych z równą liczebnością grup badanych. Należy pamiętać, że przy nierównych liczebnościach grup sumaryczna liczebność jest zawsze wyższa niż podwojona liczebność jednej grupy.

Należy też pamiętać, że obliczana liczebność jest wartością szacunkową. Innymi słowy, pomaga ona wskazać, czy liczebność powinna być na przykład bliższa 50 czy 100 obserwacji, ale nie rozróżnia między 49 a 51. Liczebność dobrze jest ocenić dla kilku różnych alternatywnych scenariuszy badawczych, a nie dla jednego określonego, gdyż takiego nie można nigdy ustalić poprawnie *a priori*. Gdybyśmy zawczasu znali odpowiedź na stawiane pytania i hipotezy, nie byłoby sensu przeprowadzać doświadczenia.

Szacowaną liczebność zwiększa się w niektórych przypadkach, na przykład przy nie całkiem losowym doborze próby, przy obecności zmiennych współtowarzyszących (*confounding variables*) czy w prospektywnych badaniach kliniczno-kontrolnych (*case-control studies*).

Występuje istotny związek między liczebnością próby a wielkością błędu II rodzaju oraz mocą testu. Ponieważ dążymy zawsze do tego, aby stosować w miarę możliwości testy o największej mocy, interesuje nas często, jaka jest moc testu przy określonej liczebności próby oraz różnicy, którą chcemy wykryć. Moc testu zależy od czterech czynników: 1) liczebności próby, 2) zmienności (rozrzutu) próby, 3) wielkości błędu I rodzaju oraz 4) wielkości wykrywanej różnicy. Możemy ją określić z równania:

$$z_{\beta} = \frac{\delta}{\sqrt{\frac{2\sigma^2}{n}}} - z_{\alpha/2}$$

gdzie:

σ^2 oznacza wariancję (zmienność) próby,

δ jest wartością różnicy, którą chcemy wykryć, jeżeli $\mu_1 \neq \mu_2$,

n jest liczebnością każdej z porównywanych grup (zakładając równe liczebności),

$z_{\alpha/2}$ jest wartością krytyczną odpowiadającą istotności, z jaką chcemy weryfikować prawdziwość hipotez.

Modyfikacja któregoś z tych czynników będzie wpływać na zmiany mocy testu. Jeżeli wykrywalna różnica wzrasta, to wzrasta także moc testu; podobnie, ze wzrostem liczebności moc testu rośnie; zwiększenie rozrzutu (zmienności) próby obniża moc testu, podobnie obniżenie prawdopodobieństwa błędu I rodzaju wiąże się z niższą mocą testu, gdyż oba rodzaje błędów są związane odwrotną proporcjonalnością. W praktyce najlepszą metodą zmniejszania obu rodzajów błędów, a więc i zwiększania mocy testu, jest zwiększenie liczebności próby. Mało liczne próby charakteryzuje niska moc testu i łatwiej jest w nich nie wykryć różnicy, gdyż wyniki testu będą statystycznie nieistotne.

Przykłady różnych zastosowań równań oceny liczebności znajdzie Czytelnik w *„Części II – Uzupełnienia, przykłady i zadania"*.

Tab. 4. Równania do obliczania minimalnej liczebności próby: (a) dla oceny istotności różnic między grupami, (*b*) dla szacowania określonej wartości z żądaną precyzją.

	musimy znać		**równanie**
(a) istotność różnic			
1. pojedyncza średnia	u, v	jak poniżej	$\frac{(u+v)^2\sigma^2}{(\mu-\mu_0)^2}$
	μ-μ_0	różnica między średnią badaną μ i średnią teoretyczną μ_0 (H_0)	
	σ	odchylenie standardowe	
2. pojedyncza częstość	μ	częstość	$\frac{(u+v)^2\mu}{(\mu-\mu_0)^2}$
	μ_0	wartość dla H_0	
	u, v	jak poniżej	
3. pojedyncza proporcja	π	proporcja	$\frac{\{u\sqrt{[\pi(1-\pi)]}+v\sqrt{[\pi_0(1-\pi_0)]}\}^2}{(\pi-\pi_0)^2}$
	π_0	wartość dla H_0	
	u, v	jak poniżej	
4. porównanie dwóch średnich (*liczebność każdej grupy*)	u, v	jak poniżej	$\frac{(u+v)^2(\sigma_1^2+\sigma_2^2)}{(\mu_1-\mu_2)^2}$
	μ_1-μ_2	różnica między średnimi	
	σ_1, σ_2	odchylenie standardowe	
5. porównanie dwóch częstości (*liczebność każdej grupy*)	u, v	jak poniżej; μ_1, μ_2 częstości	$\frac{(u+v)^2(\mu_1+\mu_2)}{(\mu_1-\mu_2)^2}$

	musimy znać		równanie
6. porównanie dwóch proporcji (*liczebność każdej grupy*)	u, v	jak poniżej; π_1, π_2 proporcja	$\frac{\{u\sqrt{[\pi_1(1-\pi_1)+\pi_2(1-\pi_2)]}+v\sqrt{[2\bar{\pi}(1-\bar{\pi})]}\}^2}{(\pi_2-\pi_1)^2}$ gdzie $\bar{\pi}=\frac{\pi_1-\pi_2}{2}$
7. badanie typu *case-control* (*liczebność każdej grupy*)	π_1	proporcja w grupie kontrolnej wystawiona na działanie czynnika	$\frac{\{u\sqrt{[\pi_1(1-\pi_1)+\pi_2(1-\pi_2)]}+v\sqrt{[2\bar{\pi}(1-\bar{\pi})]}\}^2}{(\pi_2-\pi_1)^2}$ gdzie $\bar{\pi}=\frac{\pi_1+\pi_2}{2}$
	OR	iloraz szans	
	π_2	proporcja w grupie badanej (wystawiona na działanie czynnika) obliczona jako:	$\pi_2=\frac{\pi_1 OR}{1+\pi_1(OR-1)}$
	u	punkt krytyczny jednostronny rozkładu normalnego odpowiadający proporcji (100%-moc testu), tzn. jeżeli moc wynosi na przykład 90%, to (100%- 90%) = 10% i u=1.28	
	v	punkt krytyczny rozkładu normalnego odpowiadający wymaganemu (obustronnemu) poziomowi istotności, np. dla istotności 5%, v=1.96	

Tab. 4. Równania do obliczania minimalnej liczebności próby: (a) dla oceny istotności różnic między grupami, (*b*) dla szacowania określonej wartości z żądaną precyzją.

	musimy znać		**równanie**
(b) precyzja			
8. pojedyncza średnia	σ	odchylenie standardowe	$\frac{\sigma^2}{e^2}$
	e	żądana wielkość błędu std.	
9. pojedyncza częstość	μ	częstość	$\frac{\mu}{e^2}$
	e	żądana wielkość błędu std.	
10. pojedyncza proporcja	π	proporcja	$\frac{\pi(1-\pi)}{e^2}$
	e	żądana wielkość błędu std.	
11. różnica między dwoma średnimi (*liczebność każdej grupy*)	σ_1, σ_2	odchylenia standardowe	$\frac{\sigma_1^2 + \sigma_2^2}{e^2}$
	e	żądana wielkość błędu std.	
12. różnica między dwoma częstościami (*liczebność każdej grupy*)	μ_1, μ_2	częstości	$\frac{\mu_1 + \mu_2}{e^2}$
	e	żądana wielkość błędu std.	
13. różnica między dwoma proporcjami (*liczebność każdej grupy*)	π_1, π_2	proporcje	$\frac{\pi_1(1-\pi_1) + \pi_2(1-\pi_2)}{e^2}$
	e	żądana wielkość błędu std.	

W przypadku estymacji liczebności dla proporcji, wymagana liczebność jest szacowana w jednostkach mianownika równania.

Rozdział 8

Transformacja danych – sposoby „normalizacji" rozkładu

W praktyce dość często zdarza się, że zebrane przez nas obserwacje nie spełniają wymagań niezbędnych dla zastosowania testów i metod, które są szczególnie użyteczne, dogodne i które lubimy używać. Tak jest na przykład ze stosowaniem testu *t* Studenta – jest on tak popularny i chętnie wykorzystywany, że najczęściej nie sprawdzamy nawet, czy nie są naruszone warunki usprawiedliwiające jego zastosowanie. Dwa z takich przeciwwskazań, którym przypisuje się największe znaczenie to normalność rozkładu – a mówiąc „normalność" myślimy *de facto* o symetryczności rozkładu – oraz jednorodność wariancji. Z transformacji korzystamy najczęściej w przypadkach lewo- lub prawoskośnych rozkładów, nierównych wariancji porównywanych prób czy nieliniowych zależności między zmiennymi. Zależnie od sytuacji i charakteru danych stosuje się różne przekształcenia matematyczne, i nie ma tutaj zbyt dużej dowolności. Inne transformacje są użyteczne w przypadku rozkładów lewoskośnych, inne dla rozkładów prawoskośnych, inne w przypadku heteroscedastyczności zmiennych, jeszcze inne w różnych wariantach nieliniowych zależności między zmiennymi. Transformacja odwrotnej proporcjonalności jest na przykład silniejsza, zaś pierwiastkowa słabsza niż logarytmiczna, i dlatego dobiera się je w zależności od stopnia skośności rozkładu. W przypadku danych procentowych lub proporcji (których rozkłady są raczej bardziej dwumianowe niż normalne, a odstępstwa od normalności są szczególnie rażące dla niskich i wysokich %, tzn. 0-30% i 70-100%) stosuje się często transformacje *arcus sinus*.

Krótki przewodnik pokazujący czym się kierować przy wyborze rodzaju transformacji matematycznej danych zawiera Tabela 5.

Transformacja logarytmiczna

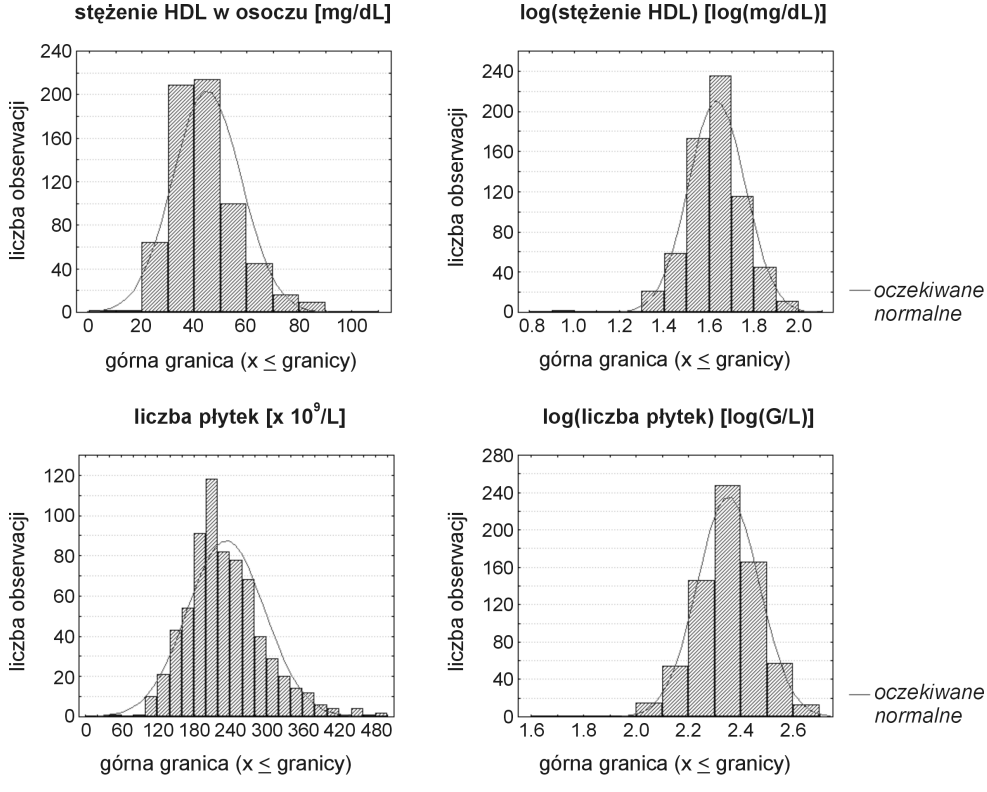
Najczęściej wykorzystywana jest transformacja logarytmiczna (z reguły wykorzystuje się logarytmy dziesiętne):

$$u = \log x$$

Oczywiście można ją stosować jedynie wtedy, gdy dane są liczbami dodatnimi i większymi od zera, ponieważ logarytmy liczb ujemnych nie istnieją, a logarytm liczby zero wynosi minus nieskończoność (istnieją wyjątkowe przypadki, które omówiono w części z przykładami). Transformacja logarytmiczna *„posiada zdolność"* rozciągania dolnej skali

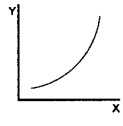
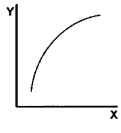
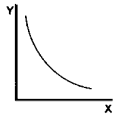
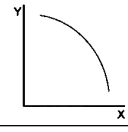
danych oraz ścieśniania górnej części (Ryc. 7). Jest szczególnie chętnie stosowana w przypadku transformacji danych o rozkładach prawoskośnych oraz gdy wariancje prób są istotnie różne (Tab. 5).

Dla danych o rozkładach silnie prawoskośnych lepszą miarą centralną jest średnia geometryczna, która jest zawsze niższa niż odpowiednia średnia arytmetyczna, gdyż nie jest tak wrażliwa na obecność danych o wysokich i bardzo wysokich wartościach z górnej części skali. Przedział ufności dla średniej geometrycznej jest oczywiście niesymetryczny względem miary centralnej (zobacz *„Część II – Uzupełnienia, przykłady i zadania"*) i dlatego nie jest miarą łatwą do interpretacji. Podobnie, odchylenia standardowe średniej geometrycznej jest multiplikatywną (iloczynopochodną), a nie addytywną miarą rozrzutu i z tych samych przyczyn nie jest często stosowane.



Ryc. 7. Rozkład danych rzeczywistych (po lewej) oraz rozkład lognormalny (po transformacji logarytmicznej) (po prawej) dla stężenia HDL w osoczu (u góry) oraz liczby płytek krwi (u dołu).

Tab. 5. Krótki przewodnik do transformacji danych.

sytuacja	**rodzaj transformacji**	
*rozkład prawoskośny (grupa **A**)*		
lognormalny	logarytmiczna ($u = log\ x$)	
bardziej skośny niż lognormalny	odwrotnej proporcjonalności ($u = 1 / x$)	
mniej skośny niż lognormalny	pierwiastkowa ($u = \sqrt{x}$)	
rozkład lewoskośny (*grupa **B***)		
umiarkowanie skośny	kwadratowa ($u = x^2$)	
bardziej skośny	sześcienna ($u = x^3$)	
nierówność wariancji		
SD proporcjonalne do średniej	logarytmiczna ($u = log\ x$)	
SD proporcjonalne do średniej^2	odwrotnej proporcjonalności ($u = 1 / x$)	
SD proporcjonalne do $\sqrt{średnia}$	pierwiastkowa ($u = \sqrt{x}$)	
dane procentowe (0-100%)	arcus sinus ($p' = \arcsin \sqrt{p}$)*	
proporcje (p, X/n; 0–1)	arcus sinus ($p' = \arcsin \sqrt{p}$)*	
	$p' = \arcsin \sqrt{\frac{X + \frac{3}{8}}{n + \frac{3}{4}}}$	
zależności nieliniowe	zmienna *y*	zmienna *x*
	grupa **A**	grupa **B**
	grupa **B**	grupa **A**
	grupa **A**	grupa **A**
	grupa **B**	grupa **B**

* $p = (\sin\ p')^2$

Odstające obserwacje

Odstającymi nazywamy nietypowe (z definicji), nie pasujące do innych, rzadko występujące obserwacje w próbie. Wierzymy, że odstające obserwacje (*outliers*, *outlying observations*) są manifestacją losowego błędu, który chcielibyśmy kontrolować i eliminować częstość obserwacji odstających i nie pasujących do ogółu. Niestety nie jest znana żadna metoda sprawdzająca się przy automatycznym usuwaniu odstających obserwacji. Dlatego też, jesteśmy zdani na analizę rozkładów pojedynczych zmiennych oraz wykresów rozrzutu dla par lub kilku zmiennych. Usuwanie zmiennych w oparciu o intuicyjne przeświadczenie ich „*inności*" może graniczyć z manipulacją danymi, dlatego staramy się dobrać jak najbardziej obiektywne metody statystyczne, i stosujemy często równolegle kilka technik weryfikacji ich „*niedopasowania*" do reszty danych. Z samej definicji odstających obserwacji wynika, że są to dane o skrajnych wartościach w monotonicznym szeregu obserwacji, obdarzone na tyle dużym błędem losowym, że nie mieszczą się w zakresie zmienności wyznaczonej przez pozostałe obserwacje próby. Ponieważ wiele czynników może być odpowiedzialnych za generowanie takich nietypowych wyników, bardzo pożądane jest zweryfikowanie przyczyn, które złożyły się na ten błąd. Próba wyeliminowania tego błędu przy powtarzaniu doświadczenia/pomiaru jest dla nas najlepszą weryfikacją występowania przypadkowości lub regularności w odstawaniu niektórych wyników. Nawet pojedyncze obserwacje różne od pozostałej zbiorowości wyników mogą istotnie zaważyć na wartości tendencji centralnej próby: mają one silny wpływ na wartość średniej i odchylenia standardowego, natomiast słabo wpływają zwykle na wartość mediany. Czym liczniejsza próba, dla której testujemy występowanie obserwacji odstających, tym wyraźniej zaznacza się takie odstawanie od reszty wyników. Dla prób o małej liczności, wykazanie i udowodnienie statystyczne, że obserwacja jest nietypowa jest trudniejsze. Najprostszą techniką łagodzenia wpływu odstających skrajnych obserwacji jest zastępowanie ich wartościami bezpośrednio przylegających danych (uporządkowanych w szeregu monotonicznym). Nie jest to właściwie test do weryfikowania, która wartość jest odstająca ze statystycznego punktu widzenia, ale ta prosta metoda umożliwia szybkie sprawdzenie, jaki efekt mają pojedyncze odstające obserwacje na miary centralne i miary rozproszenia próby. Istnieje wiele „*jednowymiarowych*" testów umożliwiających statystyczną weryfikację, czy określona obserwacja może być uznana za odstającą i odrzucona; do najczęściej stosowanych należą test Dixona, test Grubbsa i reguła „*czterech sigma*" (4σ). Wszystkie one opierają się na założeniu normalności rozkładu, a ich moc zależy od liczebności próby. W przypadku metod dwu- lub wielowymiarowych decyzja o odrzucaniu odstających obserwacji opiera na rachunku rozkładu reszt między wartościami obserwowanymi i przewidywanymi (odstawaniem wartości od „*teoretycznej*" linii regresji):

$$\sum \left(y_i - \overline{y}\right)^2 = \sum \left(y_c - \overline{y}\right)^2 + \sum \left(y_i - \overline{y}\right)^2$$

$$\text{czyli } SS_{\text{całkowita}} = SS_{\text{wyjaśniona}} + SS_{\text{niewyjaśniona}}$$

oraz szacowaniu tzw. „*studentyzowanego*" rozkładu zmienności niewyjaśnionej (reszt):

$$t = \frac{y_i - y_c}{\sqrt{MS_{błędu}}}.$$

Zastosowanie tych metod omówiono dokładniej w „*Części II – Uzupełnienia, przykłady i zadania*".

Rozdział 9

Ustalanie relacji między zmiennymi: zależności statystyczne i ciągi przyczynowo-skutkowe

Zasadniczym i ostatecznym celem większości analiz i testów statystycznych wykorzystywanych do badania zależności między zmiennymi jest ocena relacji zachodzących między zmiennymi. Metody te pozwalają nam określić charakter i siłę takich relacji, ale nie odpowiadają na pytanie, co jest przyczyną a co skutkiem w badanym procesie/zjawisku. Na podstawie określenia takiej relacji badacz może – opierając się na zdobywanym doświadczeniu i korzystając z gromadzonej wiedzy – podjąć próbę bliższej charakterystyki interesujących zjawisk czy procesów. Większość takich metod opiera się na tym, że ocenia stosunek pewnej wspólnej zmienności badanych przez nas zmiennych do ich ogólnej (całkowitej) zmienności. Może to być na przykład stosunek tej części całkowitej zmienności reaktywności płytek krwi, którą można wyjaśnić wpływem lipoprotein LDL (jest to tzw. zmienność wyjaśniana) do całkowitej zmienności reaktywności płytek. Zmienność wyjaśniana oznacza tę część zmienności jednej zmiennej, która może być wyjaśniona wartościami drugiej zmiennej i odwrotnie.

Ustalając relację między zmiennymi powinniśmy mieć od początku świadomość jaki jest zasadniczy cel naszej analizy, ponieważ ten wstępny wybór jest krytyczny przy naszym dalszym zainteresowaniu jedną z dwóch podstawowych metod badania zależności między zmiennymi: metodą regresji lub metodą korelacji. Podczas gdy pierwsza z tych metod jest wykorzystywana do przewidywania (predykcji) wartości jednej zmiennej na podstawie wartości drugiej, druga służy do wyrażania jak bardzo zmiany jednej zmiennej korespondują ze zmianami drugiej. To rozgraniczenie jest często nie przestrzegane wśród niedoświadczonych badaczy, a wynika to w dużej mierze z faktu, że od strony rachunkowej oba typy analizy są bardzo podobne, zaś estymowane parametry mogą być wzajemnie przeliczalne. Optymalnie, do właściwego rozstrzygania o tym, która z metod nadaje sie lepiej do analizy badanego procesu czy zjawiska, wykorzystujemy naszą teoretyczną wiedzę na temat charakteru badanej zależności. Ogólnie, w przypadku zależności o charakterze liniowym, metodę korelacji stosujemy do badania wzajemnych zależności zmiennych zależnych, zaś metodę regresji do badania zależności zmiennej (zmiennych) zależnej od zmiennej (zmiennych) niezależnych (manipulowalnych przez badacza).

Korelacja liniowa

Korelacja liniowa (Pearsona) (*linear/Pearson's correlation*) jest typem analizy korelacyjnej stosowanej do badania zależności statystycznej zmiennych zależnych o charakterze ciągłym. Zmienne te powinny być zatem wyrażone w odpowiedniej skali, np. przedziałowej lub ilorazowej. Inne wymagania zastosowania tej metody to oczywiście typowe dla testów parametrycznych założenie o normalności rozkładu. Siłę korelacji określamy podając wartości tzw. współczynnika korelacji (*correlation coefficient*):

$$r = \frac{\sum \left(x - \bar{x}\right)\left(y - \bar{y}\right)}{\sqrt{\left[\sum \left(x - \bar{x}\right)^2 \sum \left(y - \bar{y}\right)^2\right]}}$$

który jest miarą zależności zmian jednego parametru ze zmianami innego.

Wartości współczynnika korelacji (r) mieszczą się w zakresie od $r = -1.0$ do $r = +1.0$. Wartość $r = -1.0$ reprezentuje idealną korelację ujemną – oznacza to, że zmiany jednego parametru odpowiadają idealnie odwrotnie proporcjonalnym zmianom drugiego parametru. Odpowiednio, wartość $r = +1.0$ oznacza doskonałą korelację dodatnią, zaś wartość $r = 0.0$ wyraża brak korelacji (Ryc. 8). Maksymalne wartości współczynnika korelacji oznaczają, że punkty leżą na prostej, czym większy rozrzut punktów, tym wartości r silniej odbiegają od –1.0 lub +1.0.

Dla wygody pod względem rachunkowym korzysta się często z następujących przekształceń wyrażeń licznika i mianownika w powyższym równaniu:

$$\sum (x - \bar{x})(y - \bar{y}) = \sum xy - \left(\sum x\right)\left(\sum y\right)/n$$

$$\sum \left(x - \bar{x}\right)^2 = \sum x^2 - \left(\sum x\right)^2/n$$

$$\sum \left(y - \bar{y}\right)^2 = \sum y^2 - \left(\sum y\right)^2/n$$

Współczynnik korelacji wyraża się wtedy wzorem:

$$r = \frac{\left[\sum xy - \left(\sum x\right)\left(\sum y\right)/n\right]}{\sqrt{\left[\sum x^2 - \left(\sum x\right)^2/n\right] * \left[\sum y^2 - \left(\sum y\right)^2/n\right]}}$$

Korelacja liniowa określa stopień wzajemnej proporcjonalności wartości dwóch zmiennych. Jeżeli przyjmujemy, że zmienne są proporcjonalne, to rozumiemy pod tym, że są zależne liniowo, czyli korelacja jest tym silniejsza, im lepiej może być zobrazowana przy pomocy linii prostej (odpowiednio nachylonej w zależności od tego czy korelacja jest dodatnia czy ujemna). Linia taka, na której – albo wokół której – rozmieszczone są punkty obrazujące graficznie zależność między zmiennymi, jest nazywana linią regresji i pod względem rachunkowym jest określana przy użyciu metody najmniejszych kwadratów. Metoda ta polega na takim dobraniu położenia w układzie współrzędnych estymowanej linii, aby suma kwadratów odległości punktów doświadczalnych od tej linii była jak najmniejsza. Fakt podnoszenia tych odległości do kwadratu sprawia, że różnice między

małymi a dużymi odległościami zostają dodatkowo „*wyolbrzymione*", a tym samym – wartość współczynnika korelacji jest bardzo czuła na sposób rozmieszczenia danych.

Oczywiście, wartość współczynnika korelacji nie zależy od jednostek miary, w jakich wyrażamy badane zmienne.

Współczynnik korelacji podniesiony do kwadratu daje tzw. ***współczynnik determinacji*** (*determination coefficient*) (r^2 lub R^2), który wyraża proporcję wspólnej zmienności dwóch (lub więcej) zmiennych. Ta wspólna zmienność jest oczywiście tym wyższa, im większa jest zależność między zmiennymi (ponieważ tym bardziej zmiany jednej zmiennej determinują zmiany drugiej zmiennej).

Ponieważ współczynnik korelacji nie jest liniową funkcją opisującą siłę relacji między zmiennymi, nie można uśredniać współczynników korelacji. Średnia wartość współczynników korelacji z wielu próbek nie będzie równa średniej korelacji w tych wszystkich próbkach. Jeśli zachodzi potrzeba oszacowania takiej „*połączonej*" („*spoolowanej*") korelacji, wówczas musimy najpierw współczynniki korelacji zamienić na inne, addytywne, mierniki, takie jak np. kwadraty współczynników korelacji – tzw. współczynniki determinacji, lub wartości *z* Fishera.

W przypadku występowania korelacji nieliniowych (*non-linear correlation*) wartość współczynnika korelacji Pearsona jest zaniżona i nie odzwierciedla prawdziwej siły związku. Lepiej jest wtedy obliczyć tzw. statystykę eta (η):

$$\eta = \sqrt{1 - \frac{\sum\left(y_i - \overline{y}_c\right)^2}{\sum\left(y_i - \overline{y}_t\right)^2}}$$

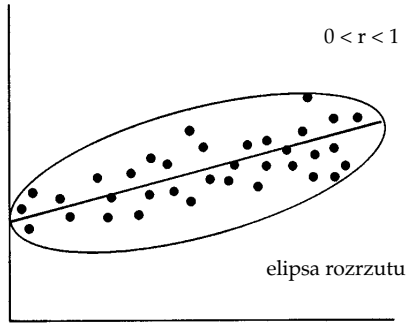
gdzie $\overline{y}_t$ jest średnią wszystkich wartości y, a $\overline{y}_c$ średnią dla kategorii (zobacz „*Część II – Uzupełnienia, przykłady i zadania*").

Istotność korelacji

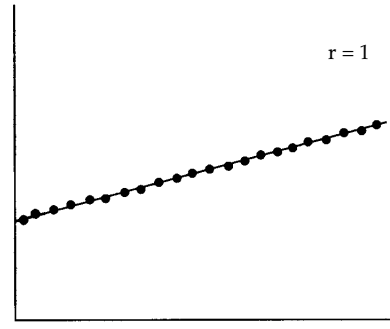
Do oceny czy obliczona przez nas wartość współczynnika korelacji jest istotnie różna od zera (czy też jest różna od zera jedynie przez czysty przypadek) wykorzystujemy test *t* Studenta. Na podstawie wartości statystyki tego testu, obliczanej jako:

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

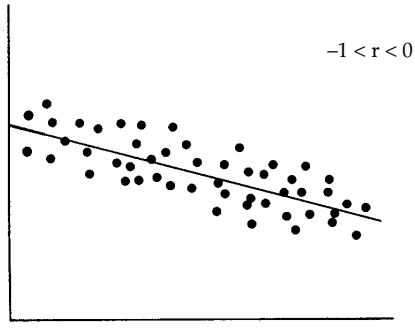
dla liczby stopni swobody $d.f. = n - 2$, szacujemy istotność korelacji (*significance of correlation coefficient*). Istotność współczynnika korelacji mówi nam o *wiarygodności* ustalonej zależności. Chociaż jest ona mniej oczywista intuicyjnie niż wartość współczynnika korelacji (jako miary siły korelacji), to jej znaczenie jest szczególnie duże w przypadku analizy prób o małych liczebnościach. Ogólnie, istotność korelacji stanowi o reprezentatywności wyniku uzyskanego na podstawie analizy pobranej losowo próby w odniesieniu do całej populacji, z której ta próba pochodziła. Wartość tej istotności mówi nam, jakie jest prawdopodobieństwo tego, że oceniona przez nas relacja zmiennych byłaby analogiczna (jak ta stwierdzona w badanej przez nas akurat próbie), gdybyśmy powtórzyli doświadczenie na innych



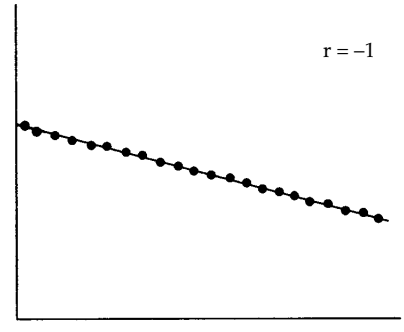
korelacja dodatnia



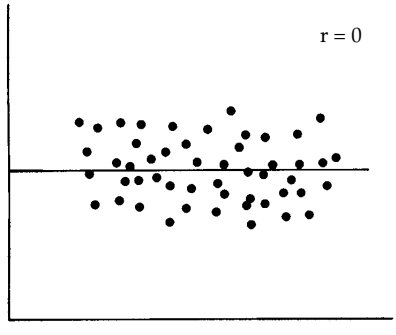
idealna korelacja dodatnia



korelacja ujemna



idealna korelacja ujemna



brak korelacji

Ryc. 8. Graficzna prezentacja różnych wartości współczynnika korelacji.

próbach pobranych losowo z tej samej populacji ogólnej. Zabiegamy o to dlatego, że wyrażając swoją opinię o sile zależności między badanymi zmiennymi traktujemy taką opinię jako określenie pewnej prawidłowości w zakresie badanego zjawiska czy procesu.

Jeszcze inaczej ujmując sens istotności korelacji, możemy powiedzieć, że istotność związku między zmiennymi precyzuje nam jak bardzo prawdopodobne jest uzyskanie obserwowanej (lub większej) siły zależności w wylosowanej przez nas próbie określonej wielkości, przy założeniu, że zależność ta nie istnieje w ogóle w populacji generalnej. W takim ujęciu, wartość poziomu istotności korelacji informuje nas o prawdopodobieństwie popełnienia błędu polegającego na odrzuceniu hipotezy, że zależność, którą badamy, faktycznie nie występuje w populacji generalnej, a jej wykrycie w badanej próbie jest jedynie dziełem przypadku.

Dla takich samych wartości współczynnika korelacji jego istotność zależy silnie od liczności próby, dla której została oszacowana. Dlaczego tak jest? Jeśli mamy do czynienia z małą liczbą obserwacji, wówczas istnieje też mała liczba wszystkich możliwych kombinacji różnych wartości poszczególnych zmiennych. Zatem prawdopodobieństwo tego, że przez przypadek wystąpi kombinacja wskazująca na silną zależność jest względnie duże. Im mniejsza jest liczność próbki, tym częstość takiego błędnego wyniku będzie większa, wskazując tym samym na istnienie zależności, która faktycznie w populacji generalnej nie występuje. Jeśli siła zależności między zmiennymi jest obiektywnie (to znaczy w populacji generalnej) mała, to trudno taką zależność wykazać, o ile próba nie jest bardzo liczna. Nawet jeżeli nasza próbka jest doskonale reprezentatywna, to wynik okaże się statystycznie nie istotny, jeśli próbka jest mała. Z drugiej strony, w przypadku kiedy zależność jest faktycznie (obiektywnie) bardzo silna, to możemy ją wykazać nawet jeśli próbka nie jest liczna. Mówiąc innymi słowy, jeżeli zależność jest obiektywnie słaba, wtedy niewielka liczba obserwacji nie wystarczy, żeby kogokolwiek przekonać o jej występowaniu; jeżeli natomiast zależność jest ewidentna, to dla większości obserwatorów będzie to wystarczająco przekonywujący argument, że ma miejsce określona prawidłowość. Jeśli zatem relacja jest silna, to będzie ona istotna nawet w małej próbie.

Test istotności współczynników korelacji jest wersją testu t Studenta, i tak jak ten ostatni jest oparty na założeniu o normalności rozkładu wartości resztowych (odległości lub odchyleń punktów doświadczalnych od linii regresji) zmiennej y, oraz o równości wariancji wartości resztowych dla wszystkich wartości zmiennej niezależnej x. Symulacja komputerowa z użyciem metody Monte Carlo wykazała jednak, że naruszenie obu tych założeń nie jest krytyczne dla prób o wystarczająco dużej liczebności. Jak dużej? Dość arbitralnie przyjmuje się, że w próbach o liczebności powyżej 50 wystąpienie poważnych nieprawidłowości z tytułu naruszenia powyższych założeń jest mało prawdopodobne, natomiast w próbach liczących powyżej 100 obserwacji założeniem o normalności nie potrzeba się praktycznie przejmować.

Najczęściej spotykane problemy przy badaniu zależności zmiennych metodą analizy korelacji

Obserwacje odstające

Odstające obserwacje, czyli takie, których wartości różnią się bardzo od wszystkich innych obserwacji w próbie, mają duży wpływ na wartość współczynnika korelacji. Wynika

to z faktu, że sposób szacowania współczynnika korelacji oraz wyznaczania odpowiadającej linii regresji opiera się na minimalizowaniu sumy kwadratów różnic (odchyleń), a nie zwykłej sumy różnic. Wierzymy, że takie odstające obserwacje są najczęściej manifestacją błędu losowego popełnianego przy zbieraniu wyników. Ponieważ mogą one w zasadniczy sposób zaniżyć albo zawyżyć rzeczywistą wartość współczynnika korelacji – szczególnie przy niewielkiej liczbie obserwacji, pożądane jest zawsze rozważenie usunięcia tych losowo odbiegających danych, np. w oparciu o wyniki testów odrzucania wyników niepewnych. Bardzo pomocna jest także graficzna prezentacja danych – ułatwia ona selekcję obserwacji odstających w naszej zbiorowości wyników.

Korelacje w grupach niejednorodnych

Z niejednorodnością wyników w badanej próbie mamy często do czynienia, gdy analizowane przez nas dane pochodzą z dwóch (lub więcej) różnych grup pomiarowych (to znaczy takich, w których wartości rozważanych zmiennych są zasadniczo różne). Graficznym odzwierciedleniem takiej szczególnej sytuacji są dwie (lub kilka) odrębne „*chmury*" wyników na wykresie rozrzutu (dwie elipsy rozrzutu). Jeżeli policzymy współczynnik korelacji dla wspólnej zbiorowości punktów w obu takich grupach, to wynikiem obliczeń może być wysoki współczynnik korelacji na skutek rozmieszczenia kilku oddzielnych grup punktów mimo, że prawdziwe wartości współczynników korelacji w każdej z grup (to znaczy gdybyśmy analizowali każdą grupę oddzielnie) są np. bliskie zera. Racjonalnym posunięciem w takim przypadku byłoby rozseparowanie grup i policzenie oddzielnych wartości współczynników korelacji (będących miarą zależności w każdej poszczególnej grupie) dla każdej z nich.

Relacje nieliniowe między zmiennymi

Ponieważ współczynnik korelacji *r* Pearsona mierzy liniową zależność między zmiennymi, jakiekolwiek odstępstwa od liniowości zależności między badanymi zmiennymi powodują wzrost sumy kwadratów odchyleń od linii regresji. Niemniej jednak, nawet nieliniowa zależność może przecież odzwierciedlać rzeczywisty i ścisły związek zmiennych. Są to kłopotliwe przypadki w praktyce badawczej, gdyż nie ma dobrego odpowiednika korelacji Pearsona dla relacji nieliniowych. Jeżeli mamy do czynienia z monotoniczną (rosnącą lub malejącą) krzywą opisującą badaną zależność, to można spróbować zastosować transformację danych (zobacz Rozdział „*Transformacja danych i odstające obserwacje – sposoby „normalizacji" rozkładu*") oraz obliczyć korelację liniową dla danych transformowanych. Można także zastosować metody nieparametryczne do oceny związku między zmiennymi, pamiętając jednak, że moc testów nieparametrycznych (zobacz „*Metody nieparametryczne*") jest o wiele słabsza niż w przypadku metod parametrycznych. Metody te są niejako z definicji niewrażliwe na efekty nieliniowe, gdyż oceniają związek między zmiennymi w oparciu o uszeregowanie danych w pewnym porządku, zależnym w sposób nieproporcjonalny od wartości zmiennych. Sprawne posługiwanie się tymi metodami wymaga jednak doświadczenia, a interpretacja wyników nie jest niekiedy prosta. Jeżeli zależy nam bardzo na zastosowaniu metody korelacji liniowej do zbiorowości punktów wykazujących relację nieliniową, to możemy także roboczo wydzielić mniejsze podgrupy (segmenty krzywej opisującej zależność), w taki sposób, aby punkty w tych mniejszych

podgrupach spełniały warunek liniowości, oraz przeprowadzić analizę korelacji dla każdej z mniejszych grup oddzielnie.

Analiza korelacji wielu zmiennych – macierze korelacji

Jeżeli mamy do czynienia z wieloma zmiennymi, kiedy tworzymy tzw. macierz korelacji (*correlation matrix*), czyli zestawienie współczynników r w różnych wariantach porównań zmiennych, powinniśmy oczekiwać, że poza zależnościami spodziewanymi odnajdziemy także istotne zależności nieoczekiwane. Przy przeprowadzaniu dużej liczby testów bezwzględna liczba istotnych korelacji przypadkowych będzie także duża (przy poziomie istotności $\alpha = 0.05$ w 5 na każde 100 porównań). Ponieważ nie istnieją żadne metody sortowania lub „*odsiewania*" prawdziwych (tzn. rzeczywistych zależności racjonalnie wytłumaczalnych) i przypadkowych korelacji, wszystkie wyniki nieprzewidziane i niezaplanowane powinniśmy traktować ze szczególną ostrożnością, a nawet nieufnością, i rozważać ich sens biologiczny pod kątem zgodności z innymi, niezależnymi wynikami. Najprostszym (choć nie najczęściej stosowanym z uwagi na koszty) sposobem weryfikacji mogłoby być powtórzenie pomiarów.

Brakujące dane w macierzy korelacji

Podstawowym i nienaruszalnym warunkiem obliczania miar zależności między zmiennymi jest istnienie określonej liczby par wyników. Konsekwencją tego oczywistego stwierdzenia jest to, że jeżeli liczebność zmiennych, których relację badamy, jest różna to liczba możliwych par będzie równa najniższej liczebności. W przypadku analizy korelacji wielu zmiennych w macierzy korelacji, musimy zabiegać o to, aby liczba brakujących danych („pustych pól" w macierzy korelacji) (*missing data in correlation matrix*) była jak najmniejsza. Wiele pakietów statystycznych przeznaczonych do analizy dużych macierzy korelacji wymaga wręcz, aby wszystkie pola były zapełnione. Istnieją zasadniczo dwie metody uzupełniania takich pustych pól: usuwanie braków danych parami lub przypadkami, albo zastępowanie pustych pól średnią. Obie nie są niestety bez wad. W większości pakietów statystycznych domyślnym sposobem usuwania brakujących danych podczas obliczania macierzy korelacji jest wykluczanie takich przypadków, w których brakuje przynajmniej jednego pomiaru dla choćby jednej zmiennej. Jest to tak zwane usuwanie przypadkami brakujących danych i tylko ta metoda zapewnia uzyskanie prawdziwej macierzy korelacji (chociaż może się zdarzyć, że bardzo ubogiej w kompletne przypadki), w której wszystkie współczynniki korelacji obliczano na podstawie identycznego zbioru danych. Ponieważ najczęściej zdarza się tak, że brakujące obserwacje są rozłożone losowo pomiędzy różne przypadki i zmienne, to sposób ten może prowadzić do znacznego zmniejszenia liczebności próby. W takich niesprzyjających okolicznościach korzystamy z alternatywnej metody polegającej na usuwaniu parami przypadków z brakującymi danymi. Metoda ta jest dość bezpieczna, jeżeli brakujące przypadki obejmują nie więcej niż 10% wszystkich obserwacji i są rozłożone równomiernie w macierzy korelacji. W przypadku „*systematycznego*" rozmieszczenia brakujących danych występuje tendencja do zafałszowania macierzy współczynników korelacji, gdyż dla różnych par zmiennych współczynniki te liczone są na podstawie różnych podzbiorów danych. Tak liczona macierz współczynników korelacji nie ma zagwarantowanej wewnętrznej zgodności i przechodniości między zmiennymi, a więc

nie jest to macierz w pełni prawdziwa. O zgodności zubożonej macierzy korelacji mówi nam porównanie średnich i odchyleń w różnych podzbiorach obserwacji (wykorzystywanych do liczenia korelacji między parami zmiennych). Brakujące dane można także zastępować średnią i postępowanie takie ma równie dużo zalet jak wad. Główną zaletą jest to, że pozwala ono na generowanie wewnętrznie spójnych macierzy wyników (prawdziwych macierzy korelacji), zaś do wad zaliczyć należy przede wszystkim to, że wprowadzenie jednakowych wartości średnich w miejscu brakujących danych sztucznie zmniejsza zmienność badanych cech. Czym więcej braków i czym więcej sztucznie wprowadzonych danych, tym bardziej „*dopisane*" dane zmniejszają zmienność parametru (cechy). Podobnie jak w przypadku usuwania danych parami, metody tej nie powinno się stosować w przypadkach, gdy brakuje więcej niż 10% obserwacji.

Korelacje pozorne

Tak jak wspomniano wyżej, im więcej analiz przeprowadzimy na określonym zbiorze danych, tym większa liczba wyników ma szansę przekroczyć ustalony poziom istotności przez czysty przypadek. Na przykład, licząc korelacje pomiędzy piętnastoma zmiennymi (105 wariantów porównań), możemy oczekiwać, że 5% z nich (tzn. jeden na każde 20) okaże się istotnych na poziomie $p \leq 0.05$ w sposób zupełnie przypadkowy, nawet jeżeli w populacji generalnej nie występuje w rzeczywistości żaden związek między tymi zmiennymi. Takie przypadkowo wykazane korelacje lub ogólnie korelacje, które powstają w wyniku wpływu innych zmiennych, nazywamy korelacjami pozornymi (*spurions correlations*). Zazwyczaj przyczyną występowania takich „niewytłumaczalnych" korelacji jest istotny wpływ innej zmiennej (lub zmiennych), której nie rejestrujemy, a której występowanie może kształtować wielkość współczynników korelacji między badanymi zmiennymi. Niestety, prawie nigdy nie wiemy, co jest tym ukrytym czynnikiem. Niektóre procedury statystyczne do analizy wielu zmiennych i porównań (to znaczy w przypadku, gdy szansa wystąpienia takich przypadkowych błędnych wyników jest większa) umożliwiają odpowiednie korygowanie wyników lub przewidują zabiegi dopasowywania w zależności od liczby porównań. Na przykład, chcąc odizolować wpływ innych zmiennych (towarzyszących) na wartość współczynnika korelacji zmiennych badanych, możemy skorzystać z metody liczenia korelacji cząstkowych, które uwzględniają (i usuwają) wpływ innych zmiennych.

Fakt występowania korelacji pozornych stawia przed badaczem szczególne wymagania co do ostrożności w ocenie niespodziewanych wyników badań. Nakazuje to też badaczowi sceptycyzm przy akceptacji oraz interpretacji trudnych do racjonalnego wytłumaczenia lub pozbawionych biologicznego sensu (według posiadanej przez nas wiedzy) zależnościach obserwowanych parametrów.

Regresja liniowa i wielokrotna

Ogólnym celem metody regresji jest badanie związków pomiędzy jedną lub wieloma zmiennymi niezależnymi (objaśniającymi) (*explaining/expalnatory variable*) a zmienną zależną (objaśnianą) (*explained variable*). To, którą zmienną nazwiemy zależną, jest bardzo istotne, gdyż w przeciwieństwie do metody korelacji różne alternatywy sparowania zmiennej

niezależnej i zależnej dają różne rozwiązania. Zależność między dwiema lub kilkoma zmiennymi opisuje równanie regresji postaci:

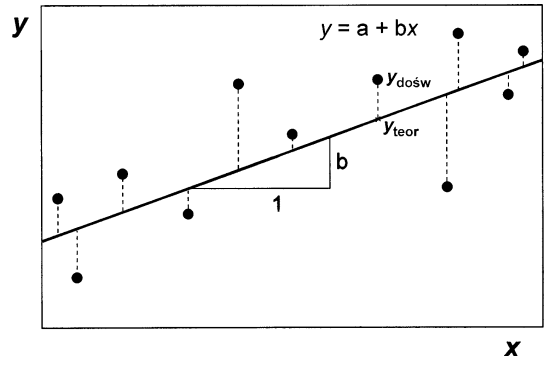
$$y = a + bx \quad \text{lub} \quad y = a_0 + a_1 x$$

W przypadku regresji liniowej (*linear regression analysis*) jest to linia prosta położona w przestrzeni dwuwymiarowej (na płaszczyźnie). Zapis równania regresji oznacza, że wartość zmiennej zależnej y może być obliczona jako suma wyrazu wolnego (a) oraz iloczynu nachylenia (b) i zmiennej niezależnej x. Wyraz wolny (a) jest nazywany stałą równania lub rzędną zerową (*intercept of regression line*), a nachylenie (b) nazywa się ***współczynnikiem regresji*** lub ***współczynnikiem kierunkowym prostej*** (*regression slope*). Jeżeli między dwiema zmiennymi nie występuje korelacja, to współczynnik b jest równy 0, a linia jest równoległa do osi odciętych układu współrzędnych.

Kierunek zależności zmiennej zależnej od zmiennej niezależnej jest zdefiniowany przez znak wartości współczynnika regresji b: wartość dodatnia oznacza relację pozytywną (kiedy wraz ze wzrostem x rośnie y), ujemna – zależność negatywną. Pod względem rachunkowym obliczanie parametrów równania regresji oraz obliczanie współczynników korelacji jest bardzo podobne; o wyborze określonej metody decyduje nasz model badawczy. Optymalne położenie linii regresji ustala się najczęściej korzystając z metody najmniejszych kwadratów, w taki sposób, aby zminimalizować sumę kwadratów wszystkich odległości punktów doświadczalnych od odpowiadających im punktów na krzywej teoretycznej ($y_{obserwowana} - y_{teoretyczna}$) (Ryc. 9). Ta suma kwadratów wszystkich odległości punktów doświadczalnych od odpowiadających im punktów na krzywej teoretycznej nazywana jest także sumą kwadratów reszt (*residual sum of squares*) lub sumą kwadratów niewyjaśnianą przez regresję. Razem z sumą kwadratów wyjaśnianą przez regresję składa się ona na całkowitą sumę kwadratów objaśniającą całkowitą zmienność parametrów, których wzajemną relację badamy. To rozbicie całkowitej sumy kwadratów na część wyjaśnianą i niewyjaśnianą (resztową) przez regresję nawiązuje do metody analizy wariancji.

zmienność	**suma kwadratów (*SS*)**	**stopnie swobody (*d.f.*)**	**błąd średniokwadratowy (*MS=SS/d.f.*)**	$F_{\alpha(1),k,v}$
regresja	$b^2 * \left[\sum x^2 - \frac{\left(\sum x\right)^2}{n}\right]$	$k = p–1$	$SS_{regr}/d.f._{regr}$	$\frac{MS_{regresji}}{MS_{resztowy}}$
resztowa	$SS_{całk} - SS_{regresji}$	$v = N - k - 1$	$SS_{reszt}/d.f._{reszt}$	
całkowita	$\sum y^2 - \frac{\left(\sum x\right)^2}{n}$	$N - 1$		

p oznacza liczbę zmiennych w modelu



Ryc. 9. W równaniu regresji, $y = a + bx$, współczynnik a jest wartością y, dla której x wynosi 0, zaś b jest przyrostem zmiennej y dla jednostkowego przyrostu zmiennej x. Wartości współczynników regresji mogą być estymowane metodą najmniejszych kwadratów w taki sposób, aby zminimalizować sumę kwadratów wszystkich odległości punktów doświadczalnych od odpowiadających im punktów na krzywej teoretycznej ($y_{obserwowana} - y_{teoretyczna}$).

W przypadkach, gdy nie występuje zależność między zmiennymi, wartość błędu średniokwadratowego regresji jest w przybliżeniu równa wartości błędu średniokwadratowego reszt (F wynosi w przybliżeniu 1). Statystyka F odpowiada testowi t do weryfikowania istotności współczynnika b lub współczynnika korelacji, r ($F = t^2$). Analiza wariancji dostarcza nam także interpretacji współczynnika korelacji: wartość ilorazu sumy kwadratów regresji do całkowitej sumy kwadratów to nic innego jak współczynnik determinacji, czyli proporcja, jaka część całkowitej zmienności jest wyjaśniana przez regresję:

$$r^2 = \frac{SS_{wyjaśniona}}{SS_{całkowita}}.$$

Współczynniki równania regresji liniowej obliczamy według wzorów:

$$b = \frac{\sum\left(x-\bar{x}\right)\left(y-\bar{y}\right)}{\sum\left(x-\bar{x}\right)^2} \quad \text{oraz} \quad a = \bar{y} - b\bar{x}$$

$$\text{lub} \quad b = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - \left(\sum x\right)^2} \quad \text{oraz} \quad a = \frac{\sum y - b\sum x}{n}.$$

Błąd współczynnika kierunkowego prostej (*error of regression slope*) oraz błąd rzędnej zerowej (*regression, intercept error*) obliczamy w następujący sposób:

$$SE_a = s\sqrt{\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum\left(x-\bar{x}\right)^2}\right]} \quad \text{oraz} \quad SE_b = \frac{s}{\sqrt{\sum\left(x-\bar{x}\right)^2}},$$

gdzie s jest odchyleniem standardowym punktów doświadczalnych od punktów teoretycznych na linii regresji, posiada n–2 stopni swobody (gdyż mamy dwa współczynniki prostej) i wynosi:

$$s = \sqrt{\left[\frac{\sum\left(y-\bar{y}\right)^2 - b^2\sum\left(x-\bar{x}\right)^2}{(n-2)}\right]}.$$

Istotność współczynnika kierunkowego (nachylenia prostej) obliczamy na podstawie porównania go z teoretyczną wartością współczynnika kierunkowego β (analogicznie do testowania istotności współczynnika korelacji):

$$t = \frac{b-\beta}{SE_b}$$

przy $d.f. = n-2$ stopniach swobody.

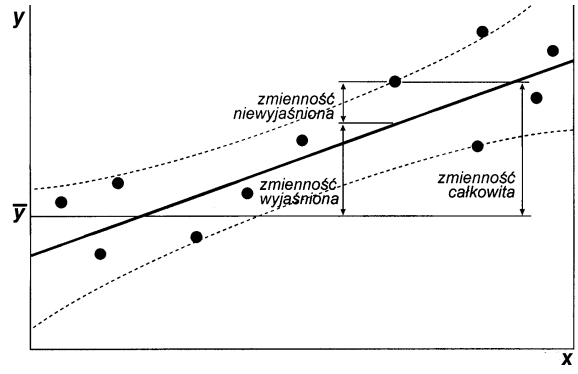
Na podstawie znajomości współczynników a i b możemy szacować (dokonywać predykcji) wartość zmiennej zależnej y dla określonych wartości zmiennej x, a także jej błąd:

$$y' = a + bx', \quad SE(y') = s\sqrt{\left[1 + \frac{1}{n} + \frac{\left(x'-\bar{x}\right)^2}{\sum\left(x-\bar{x}\right)^2}\right]}$$

Wartość błędu standardowego y' jest mniejsza dla x' bliskich średniej (Ryc. 10); ogólnie należy przestrzec przed estymacją wartości zmiennej zależnej dla wartości spoza zakresu szacowania współczynników regresji, ponieważ równanie regresji nie musi być prawdziwe (liniowe) poza zakresem estymacji współczynników.

Estymowana linia regresji wyraża najlepszą (z możliwych do ustalenia) predykcję zmiennej zależnej y dla danych wartości zmiennej niezależnej x. Dopasowanie takie nigdy nie jest idealne i zazwyczaj mamy do czynienia z mniejszymi lub większymi odchyleniami danych pomiarowych od punktów teoretycznych (obliczonych) linii regresji (odchylenia takie nazywają się *wartościami resztowymi* (*residuals*)).

Im mniejsza jest wariancja (zmienność, rozproszenie) wartości resztowych od linii regresji w stosunku do zmienności ogólnej, tym dokładniejsze i bardziej wiarygodne jest przewidywanie wartości zmiennej zależnej na podstawie zmiennej niezależnej. W przypadku braku zależności między zmiennymi, taka proporcja zmienności resztowej y do zmienności całkowitej wynosi 1, natomiast dla zmiennych idealnie zależnych od siebie proporcja taka wynosi 0, gdyż zmienność resztowa równa się 0. W praktyce wartość tej proporcji oscyluje między 0 i 1. Dopełnieniem takiej proporcji do jedności jest tzw. współczynnik determinacji (R^2), określający jaka część zmienności zmiennej zależnej y jest tłumaczone przez regresję. Wynika stąd, że czym wyższy współczynnik determinacji, tym lepiej zmienność regresji tłumaczy całkowitą zmienność y, czyli tym lepsza determinacja zmiennej zależnej (y) przez zmienną niezależną (x). Współczynnik ten stanowi zatem użyteczny wskaźnik jakości dopasowania modelu do danych, bowiem wysokie wartości R^2 wskazują, że prawie cała zmienność zmiennej zależnej może być objaśniona przez zmienną niezależną w analizowanym modelu.



Ryc. 10. Zmienność wyjaśniona i niewyjaśniona regresji oraz przedziały ufności krzywej regresji.

Założenia i ograniczenia w analizie regresji liniowej – analiza zmiennych resztowych

Dwa podstawowe założenia to: założenie normalności i założenie liniowości. Pierwsze polega na tym, że zmienne resztowe (tzn. różnice między wartościami doświadczalnymi/zmierzonymi a obliczonymi/teoretycznymi z równania regresji) podlegają rozkładowi normalnemu. Pomimo, że większość testów (np. test *F*) jest mało wrażliwa na odstępstwa od tego założenia, zwyczajowo przed wyciągnięciem ostatecznych wniosków sprawdza się jakie są rozkłady badanych zmiennych. W przypadku niespełniania normalności można stosować np. odpowiednie transformacje danych.

Założenie liniowości jest w praktyce bardzo trudne do udowodnienia, gdyż mało która relacja między parametrami spełnia założenie liniowości w pełnym zakresie wartości zmiennych. Bardziej poważne odstępstwa od tych założeń, takie jak na przykład odstające obserwacje, mogą w istotny sposób wpływać na wartości współczynników równania regresji poprzez naciąganie linii regresji w określonym kierunku. Nawet zmiana współrzędnych (lub usunięcie) pojedynczych danych może prowadzić do zupełnie różnych wyników analizy. Powszechnie stosowaną metodą identyfikacji odstępstw od założeń poprawnej analizy regresji jest analiza zmiennych resztowych (*examination of residuals*).

Ograniczenia metody regresji są podobne do tych, które spotykamy w analizie korelacji: przy pomocy metod regresji można przekonać się o istnieniu relacji oraz podać jej charakterystykę, ale nie da się dowieść istnienia związku przyczynowego będącego podłożem tej relacji.

Porównywanie dwóch lub więcej równań regresji

Badając podobieństwa między kilkoma równaniami regresji mamy do wyboru testowanie różnic współczynników kierunkowych *b* oraz współczynników *a* (rzędnych zerowych). Testy służące do tego nazywa się ogólnie testami równoległości prostych regresji. Warianty algorytmów postępowania dla różnej liczby linii regresji przedstawia poniższy schemat.

Równoległość prostych regresji bada się z wykorzystaniem testów opartych na logice testu *t* Studenta. Najprostszym wariantem takich metod jest zastosowanie testu *t* Studenta oraz obliczenie statystyki testu dla porównania dwóch prostych regresji:

$$t = \frac{b_1 - b_2}{SE_{b_1 - b_2}},$$

gdzie wartość błędu standardowego różnicy między współczynnikami wynosi:

$$SE_{b_1 - b_2} = \sqrt{\frac{\left(s^2_{x,y}\right)_p}{\sum x_1^2} + \frac{\left(s^2_{x,y}\right)_p}{\sum x_2^2}},$$

a $\left(s^2_{x,y}\right)_p$ jest *„połączoną"* (*„spoolowaną"*) zmiennością (*pooled variability*) obliczaną jako:

$$\left(s^2_{x,y}\right)_p = \frac{(SS_{reszt})_1 + (SS_{reszt})_2}{(df_{reszt})_1 + (df_{reszt})_2}.$$

Jeżeli odrzucimy hipotezę zakładającą równość współczynników kierunkowych *b,* to możemy obliczyć współrzędne punktu, gdzie dwie linie regresji przecinają się. Będzie on miał współrzędne:

$$x_{przecięcia} = \frac{a_2 - a_1}{b_1 - b_2}, \quad \text{oraz}$$

$$y_{przecięcia} = a_1 + b_1 x_{przecięcia} \quad \text{lub} \quad y_{przecięcia} = a_2 + b_2 x_{przecięcia}.$$

Jeżeli hipoteza zakładająca równość współczynników kierunkowych *b* nie zostanie odrzucona, tzn. linie są równoległe, to możemy obliczyć tzw. *„wspólny"* (ważony) współczynnik kierunkowy b_c (*regression common slope*):

$$b_c = \frac{\left(\sum xy\right)_1 + \left(\sum xy\right)_2}{\left(\sum x^2\right)_1 + \left(\sum x^2\right)_2}.$$

Testowanie równości współczynników *a* przeprowadzamy posługując się równaniami:

$$t = \frac{\left(\bar{y}_1 - \bar{y}_2\right) - b_c(\bar{x}_1 - \bar{x}_2)}{\sqrt{\left(s^2_{x,y}\right)_c \left[\frac{1}{n_1} + \frac{1}{n_2} + \frac{\left(\bar{x}_1 - \bar{x}_2\right)^2}{A_c}\right]}}$$

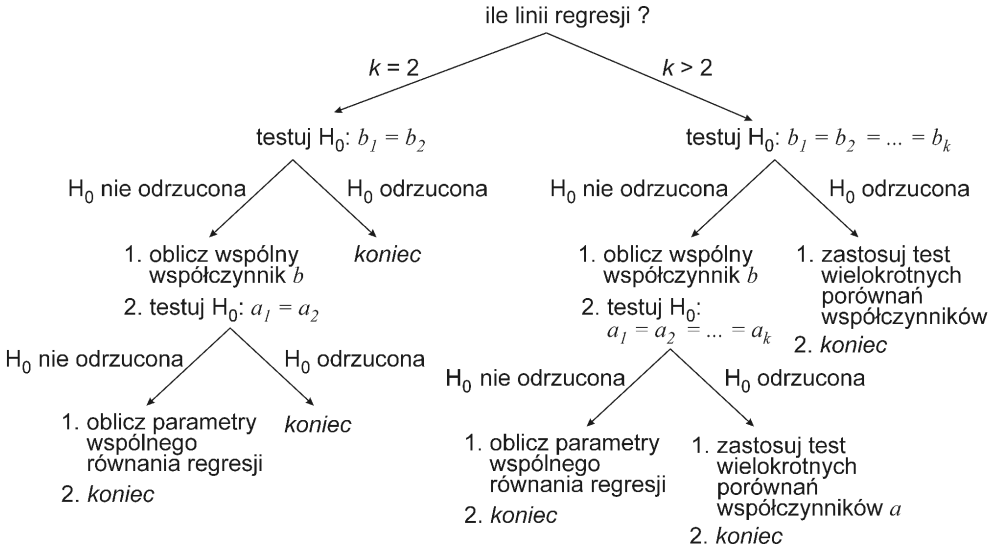
gdzie:

$$\left(s^2_{x,y}\right)_c = \frac{SS_c}{df_c} \qquad SS_c = C_c - \frac{B_c^2}{A_c} \qquad df_c = n_1 + n_2 - 3$$

$$A_c = \left(\sum x^2\right)_1 + \left(\sum x^2\right)_2$$

$$B_c = \left(\sum xy\right)_1 + \left(\sum xy\right)_2$$

$$C_c = \left(\sum y^2\right)_1 + \left(\sum y^2\right)_2$$



Dlaczego nie możemy w tej sytuacji zastosować sposobu liczenia analogicznego do tego używanego przy testowaniu równości współczynników b, to znaczy wykorzystać równania:

$$t = \frac{a_1 - a_2}{SE_{a_1 - a_2}}$$

gdzie:

$$SE_{a_1-a_2} = \sqrt{\left(s^2_{x,y}\right)_p \left[\frac{1}{n_1} + \frac{1}{n_2} + \frac{x_1^{-2}}{\left(\sum x^2\right)_1} + \frac{x_2^{-2}}{\left(\sum x^2\right)_2}\right]} \quad ?$$

Głównie dlatego, że ten ostatni test testuje punkt a, który może leżeć daleko poza zakresem wartości zmiennych, dla którego estymowano równania regresji. Punkt ten jest najczęściej jedynie ekstrapolacją hipotetycznej krzywej regresji na oś Y przy założeniu, że krzywa ta ma przebieg liniowy na całej swojej długości. W rzeczywistości jest to dużym uproszczeniem, gdyż rzadko kiedy tak jest naprawdę (Ryc. 11). Skoro wartości y dla $x = 0$ (czyli współczynniki a lub rzędne zerowe) są położone daleko od wartości wyznaczających średnie x, stąd błędy standardowe takiej estymacji będą bardzo duże, a moc testu bardzo niska.

Jeżeli nie odrzucimy hipotezy zerowej zakładającej równość współczynników a, tzn., jeżeli linie regresji będą miały nieistotnie różne rzędne zerowe, to możemy obliczyć tzw. *„wspólny"* (ważony) współczynnik a_c (*regression common intercept*):

$$a_c = \bar{y}_p - b_c \bar{x}_p$$

gdzie

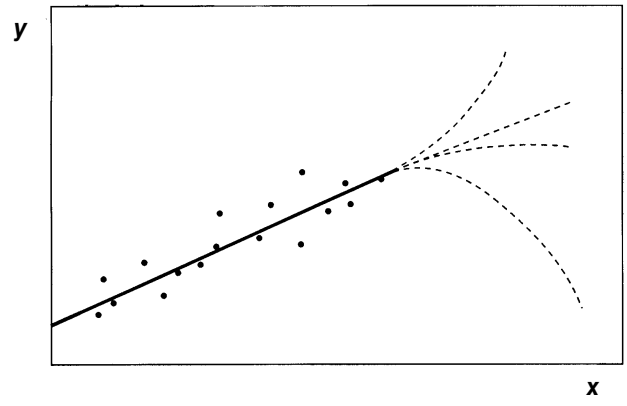
$$\bar{x}_p = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2} \quad \text{oraz} \quad \bar{y}_p = \frac{n_1\bar{y}_1 + n_2\bar{y}_2}{n_1 + n_2}$$

Oczywiście, jeżeli nie wykażemy, że współczynniki kierunkowe *b* są różne oraz że współczynniki *a* są różne, możemy zapisać wspólne równanie regresji dla wszystkich punktów obu populacji danych jako:

$$y_i = a_c + b_c x_i$$

Przy porównywaniu więcej niż dwóch współczynników kierunkowych *b* testujemy hipotezę postaci H_0: $b_1 = b_2 = ... = b_k$. Procedura obliczeń jest dość pracochłonna i polega ogólnie na oszacowaniu zmienności dla każdej linii regresji, regresji wspólnej, połączonej oraz całkowitej.

	$\sum x^2$	$\sum xy$	$\sum y^2$	SS_{reszt}	df_{reszt}
regresja 1	A_1	B_1	C_1	$SS_1 = C_1 - \frac{B_1^2}{A_1}$	$df_1 = n_1 - 2$
regresja 2	A_2	B_2	C_2	$SS_2 = C_2 - \frac{B_2^2}{A_2}$	$df_2 = n_2 - 2$
.	.	.	.	.	.
.	.	.	.	.	.
regresja *k*	A_k	B_k	C_k	$SS_k = C_k - \frac{B_k^2}{A_k}$	$df_k = n_k - 2$
regresja „połączona"				$SS_p = \sum_{i=1}^{k} SS_i$	$df_p = \sum_{i=1}^{k}(n_i - 2)$
					$= \sum_{i=1}^{k} n - 2k$
regresja „wspólna"	$A_c = \sum_{i=1}^{k} A_i$	$B_c = \sum_{i=1}^{k} B_i$	$C_c = \sum_{i=1}^{k} C_i$	$SS_c = C_c - \frac{B_c^2}{A_c}$	$df_c = \sum_{i=1}^{k} n_i - k - 1$
regresja całkowita	A_t	B_t	C_t	$SS_t = C_t - \frac{B_t^2}{A_t}$	$df_t = \sum_{i=1}^{k} n_i - 2$



Ryc. 11. Problemy ekstrapolacji poza zakresem estymacji równania regresji.

Wartość statystyki F dla porównania współczynników kierunkowych b wynosi:

$$F = \frac{\left(\frac{SS_c - SS_p}{k-1}\right)}{\frac{SS_p}{df_p}}$$

z liczbą stopni swobody dla licznika i mianownika równą odpowiednio k–1 oraz df_p.
Wartość statystyki F dla porównania współczynników a wynosi:

$$F = \frac{\left(\frac{SS_t - SS_c}{k-1}\right)}{\frac{SS_c}{df_c}}$$

z liczbą stopni swobody dla licznika i mianownika równą odpowiednio k–1 oraz df_c.
Jeżeli odrzucimy hipotezę zerową zakładającą równość wszystkich porównywanych współczynników, to możemy w dalszym ciągu zastosować któryś z testów porównań wielokrotnych, np. test Tukeya. Istotność różnic między poszczególnymi weryfikujemy obliczając wartość statystyki q:

$$q = \frac{b_B - b_A}{SE}$$

$$\text{gdzie} \quad SE = \sqrt{\frac{\left(s^2_{x,y}\right)_p}{2}\left[\frac{1}{\left(\sum x^2\right)_A} + \frac{1}{\left(\sum x^2\right)_B}\right]}.$$

Podobnie, wartość statystyki q dla porównań wielokrotnych między współczynnikami a wyliczamy z równania:

$$q = \frac{\left|\left(\bar{y}_A - \bar{y}_B\right) - b_C\left(\bar{x}_A - \bar{x}_B\right)\right|}{SE}$$

gdzie $$SE = \sqrt{\frac{\left(s^2_{x,y}\right)_C}{2}\left[\frac{1}{n_A} + \frac{1}{n_B} + \frac{\left(\bar{x}_A - \bar{x}_B\right)^2}{\left(x^2\right)_A + \left(x^2\right)_B}\right]}.$$

Możemy także zastosować ogólny całościowy test weryfikujący czy istnieje koincydencja między k równaniami regresji, tzn. sprawdzić czy wszystkie współczynniki b są sobie równe oraz czy wszystkie współczynniki a są sobie równe:

$$F = \frac{\left(\frac{SS_t - SS_p}{2(k-1)}\right)}{\frac{SS_p}{df_p}}$$

z liczbą stopni swobody dla licznika i mianownika, równą odpowiednio 2(k–1) oraz df_p.

Jeżeli nie odrzucimy hipotez zerowych H_0: $b_1 = b_2 = ... = = b_k$ oraz H_0: $a_1 = a_2 = ... = a_2$, to możemy obliczyć wartość wspólnych uśrednionych (ważonych) współczynników b i a, które charakteryzują jedną linię regresji opisującą wszystkie analizowane punkty doświadczalne (ze wszystkich grup).

Przykłady zastosowań opisanych wyżej procedur testowania równoległości linii regresji znajdzie Czytelnik w *„Części II – Uzupełnienia, przykłady i zadania"*.

Regresja wielokrotna i korelacje cząstkowe

Model regresji wielokrotnej

Z regresją wielokrotną (*multiple regression*) mamy do czynienia bardzo często w praktyce badawczej, zwłaszcza w złożonych analizach łącznego wpływu parametrów klinicznych i diagnostycznych, czy biochemicznych. Ponieważ w przypadku regresji wielokrotnej mamy do czynienia z więcej niż jedną zmienną niezależną, linii regresji nie możemy przedstawić w prosty sposób w przestrzeni dwuwymiarowej. Zależność między zmienną zależną a kilkoma zmiennymi niezależnymi możemy zobrazować równaniem:

$$y = a + b_1 * x_1 + b_2 * x_2 + ... + b_p * x_p$$

gdzie b_1, b_2,... b_p oznaczają współczynniki regresji zmiennych (parametrów) x_1, x_2,... x_p. Współczynniki te reprezentują niezależne przyczynki każdej ze zmiennych niezależnych do predykcji (przewidywania) badanej zmiennej zależnej. Współczynniki kierunkowe b_1 i b_2 są cząstkowymi współczynnikami regresji i mają sens statystyczny podobny do cząstkowych współczynników korelacji (*zobacz niżej*). Różnią się one od odpowiadających im liniowych współczynników regresji prostej liczonych przy założeniu braku interakcji między zmiennymi niezależnymi. Aby to zaobserwować przeanalizujmy przykład 67 w *„Części II – Uzupełnienia, przykłady i zadania"*.

Model regresji wielokrotnej możemy stosować dla dowolnej liczby zmiennych niezależnych, ale w miarę wzrostu tej liczby interpretacja wyników staje się coraz bardziej złożona. Niektóre parametry o charakterze czynników stałych (*fixed*) są na przykład kluczowe w każdej analizie danych klinicznych, i należy je zawsze uwzględniać lub skorygować inne zmienne włączane do modelu na ich obecność. Swoją decyzję na temat tego, które zmienne włączyć, a które nie, możemy oprzeć o wyniki badania zależności statystycznych poszczególnych zmiennych niezależnych ze zmienną zależną. Właściwego doboru zmiennych do modelu regresji wielokrotnej możemy dokonać stosując metody regresji krokowej postępującej (włączania zmiennych, *forward stepwise*) lub wstecznej (usuwanie zmiennych, *backward stepwise*).

W pierwszym przypadku badamy na początku parametry metodą prostej regresji dla każdej z rozważanych zmiennych. Parametr, dla którego zanotowaliśmy największy udział zmienności wyjaśnionej (przez regresję), włączamy jako pierwszy do modelu. Ten największy udział oznacza, że ta zmienna w największym stopniu wpływa na dopasowanie modelu. W taki sam sposób postępujemy dalej, testując (metodą regresji wielokrotnej) jak włączenie każdej ze zmiennych jako drugiej, trzeciej, ...*n*-tej wpływa na dopasowanie modelu. Włączanie kolejnych zmiennych do modelu przerywamy, gdy dodanie kolejnej zmiennej nie poprawia istotnie dopasowania modelu lub kiedy osiągamy założony wstępnie limit liczby zmiennych w modelu (*zobacz niżej*).

W drugim podejściu włączamy do modelu wszystkie dostępne zmienne (pozostając w granicach ustalonego wstępnie limitu liczby zmiennych). Następnie usuwamy kolejne zmienne z modelu sprawdzając, jak usunięcie każdej zmiennej wpływa na dopasowanie modelu regresji do danych doświadczalnych. Procedura kończy się, gdy udział wszystkich zmiennych w modelu jest istotny lub gdy osiągniemy założony limit liczby zmiennych (*zobacz niżej*).

Dla dużych liczebności (liczby przypadków) i dużej liczby zmiennych procedury te mogą być bardzo złożone pod względem rachunkowym. Dlatego też do estymacji takich stosujemy komputerowe pakiety statystyczne, w których zastosowanie algorytmów obu tych technik jest zautomatyzowane. Często zdarza się, że każda z technik „*krokowych*" prowadzi do odmiennych wyników końcowych, to znaczy lista włączonych do modelu zmiennych może się różnić w metodzie postępującego włączania i postępującego usuwania zmiennych. Oznacza to, że żadna z tych technik nie daje bezwzględnie najlepszego wyboru modelu dla określonej liczby analizowanych zmiennych. Alternatywnym rozwiązaniem jest badanie dopasowania z uwzględnianiem wszystkich możliwych kombinacji zmiennych, tzn. sprawdzenie, która pojedyncza zmienna daje najlepsze dopasowanie, która para zmiennych daje najlepsze dopasowanie, następnie która trójka daje najlepsze dopasowanie, i tak dalej. Niestety, jak zauważymy, niekoniecznie najlepiej dopasowana para zmiennych będzie uwzględniała najlepiej dopasowaną pojedynczą zmienną.

Pamiętając o niebezpieczeństwie korelacji pozornych, musimy zauważyć, że także w metodzie regresji wielokrotnej wprowadzenie wielu zmiennych do modelu może być ryzykowne, ponieważ zgodnie z zasadą losowości, prawdopodobieństwo tego, że wśród wielu elementów wielu zmiennych jakaś zmienna przez przypadek okaże się istotna, jest wcale wysokie, szczególnie w sytuacjach, gdy mamy do czynienia z niewielką liczbą obserwacji. Ogólnie zaleca się włączać do analiz wieloparametrowych przynajmniej 20 razy więcej przypadków (obserwacji, pomiarów, przypadków, pacjentów) niż występuje w niej zmiennych (parametrów badanych). W przeciwnym wypadku, oceny linii regresji będą bardzo niestabilne i będą silnie zależeć od liczby przypadków.

W przypadkach, kiedy mamy do czynienia z wieloma zmiennymi niezależnymi, które są powiązane ze sobą i silnie wzajemnie od siebie zależą (są współliniowe) może wystąpić tzw. problem złego uwarunkowania macierzy, spowodowany występowaniem *zmiennych nadmiarowych* (*redundant variables*). Nadmiarową nazywamy taką zmienną niezależną, której wpływ na zmienną zależną tłumaczą całkowicie inne zmienne w modelu.

Jeżeli mamy do czynienia z badaniem zależności wśród osobników, których możemy zaklasyfikować do (kilku) różnych grup, pożądane jest wprowadzenie do analizy regresji wielokrotnej zmiennych niezależnych dyskretnych (nieciągłych) (*multiple regression with dummy variables, with discrete explanatory variables,*); możemy w ten sposób uzyskać dodatkowe cenne informacje dotyczące np. wpływu występowania choroby, infekcji, polimorfizmu genetycznego, przedziału wiekowego, płci, itd. na zależności zmiennych.

W przypadkach, gdy jedna lub kilka analizowanych zmiennych niezależnych ma charakter nieciągły, możemy zastosować kilka sposobów. Powszechnie przyjętym rozwiązaniem jest wydzielenie kilku regionów (zwykle 3-5, zależnie od wielkości próby) zmiennej i włączenie do analizy nowej zmiennej grupującej (dyskretnej) zamiast zmiennej nieliniowej (tak jak to opisano powyżej). Wstępnie można także włączyć oba „*warianty*" zmiennej nieliniowej – dyskretny i ciągły i sprawdzić, jaki jest rzeczywisty wpływ nieliniowości na zakres zmienności. Alternatywnie, można także dokonać transformacji danych w celu linearyzacji zmiennej nieliniowej (zobacz Rozdział „*Transformacje danych i odstające obserwacje – sposoby „normalizacji" rozkładu*").

Korelacje cząstkowe

Badając zależność zmiennej zależnej od grupy zmiennych niezależnych zakładamy, że każda zmienna niezależna x_i jest skorelowana ze zmienną y przy uwzględnieniu wpływu wszystkich innych zmiennych niezależnych. Każda z rozpatrywanych zmiennych niezależnych może potencjalnie wpływać na siłę związku innych zmiennych niezależnych ze zmienną zależną. W sytuacji, gdy chcemy zbadać jaki jest wpływ takich interakcji, korzystamy z metody obliczania korelacji cząstkowej (*partial correlation*), która uwzględnia wpływ wielu towarzyszących parametrów na wielkość relacji dwóch spośród nich.

Dzięki uwzględnieniu takich interakcji dowiadujemy się jaki jest cząstkowy udział każdego z wielu badanych parametrów na wyjaśnioną zmienność innego zależnego parametru. Innymi słowy, korelacja cząstkowa charakteryzuje nam związek między dwiema badanymi zmiennymi ciągłymi przy usunięciu wpływu jakichkolwiek innych zmiennych ciągłych. Najprostszym wariantem analizy korelacji cząstkowej jest wyłączenie wspólnego efektu jakiejś zmiennej (*c*) ze związku między dwiema innymi badanymi zmiennymi (*a*,*b*). Tak szacowany współczynnik korelacji cząstkowej wynosi:

$$r_{ab,c} = \frac{r_{ab} - (r_{ac})(r_{bc})}{\sqrt{\left(1 - r_{ac}^2\right)\left(1 - r_{bc}^2\right)}}$$

gdzie $r_{ab,c}$ oznacza współczynnik korelacji między zmiennymi *a* i *b* z wyłączeniem wpływu zmiennej *c* (zobacz też „*Część II – Uzupełnienia, przykłady i zadania*").

Podsumowanie

- Korelacja i regresja pozwalają na szacowanie związków między zmiennymi, a nie służą do wyznaczania ciągu przyczynowo-skutkowego. Nie można dowieść istnienia związku przyczynowego opierając się wyłącznie na współczynnikach korelacji czy regresji. Bez naszej wiedzy o zjawisku czy procesie nie możemy wiedząc o występowaniu zależności i znając siłę takiej relacji powiedzieć, co jest przyczyną (czynnikiem sprawczym), a co skutkiem (efektem działania czynnika).
- Korelacja służy zasadniczo do badania zależności zmiennych zależnych; do badania zależności między zmienna zależną a zmienną niezależną stosujemy metodę regresji.
- Wraz ze zwiększaniem liczby analizowanych przypadków w wylosowanej próbie, wzrasta prawdopodobieństwo korelacji pozornych, czyli wzrasta liczba wyników, które mają szansę przekroczyć ustalony poziom istotności przez czysty przypadek. Stawia przed badaczem wymóg szczególnej ostrożności w ocenie niespodziewanych, trudnych do racjonalnego wytłumaczenia, lub pozbawionych biologicznego sensu zależności obserwowanych parametrów.
- Dobierając rodzaj krzywej do punktów doświadczalnych kierujemy się przede wszystkim wiedzą o zachodzeniu procesu/zjawiska, a nie rozkładem danych doświadczalnych, który w dużej mierze może być wynikiem błędów systematycznych pojawiających się podczas każdego badania.
- Poziomy istotności współczynnika korelacji oraz współczynników regresji są podstawowym źródłem informacji o wiarygodności oszacowanej miary zależności zmiennych.
- Mała liczebność próby utrudnia, a niekiedy całkowicie uniemożliwia, wiarygodne orzekanie o występowaniu istotnej zależności między zmiennymi. W mało licznych próbach trudniej wykazać wiarygodność badanej zależności.
- Nie jest dopuszczalne uśrednianie współczynników korelacji dla kilku prób należących do jednej większej populacji; stosujemy wtedy inne zabiegi liczenia miary zależności reprezentatywnej dla większej grupy powstałej z fuzji kilku mniejszych prób.
- Graficzna analiza wykresów rozrzutu jest niezbędnym elementem w metodzie regresji i metodzie korelacji; powinna ona być jedną ze wstępnych czynności zmierzających do właściwej oceny zależności między zmiennymi.
- Niejednorodność danych może być istotnym czynnikiem wpływającym na wartość szacowanego współczynnika korelacji oraz współczynników krzywych regresji.
- W analizach wieloparametrowych, takich jak np. metoda regresji wielokrotnej, powinniśmy uwzględniać przynajmniej 20 razy więcej przypadków (obserwacji, pomiarów, przypadków, pacjentów) niż występuje w niej zmiennych (parametrów badanych). W przeciwnym wypadku oceny linii regresji będą bardzo niestabilne i będą silnie zależeć od liczby przypadków.

Rozdział 10

Tabele liczebności i statystyki oparte na charakterystyce testu chi^2

W przypadku, gdy badamy zmiany liczebności oraz zmiany ich wzajemnych proporcji stosujemy tak zwane wielodzielcze (wielopolowe) tabele liczebności (*contingency tables*). Są to – najogólniej mówiąc – tabele krzyżowe z dwoma lub więcej czynnikami, które służą nam do badania zależności między zmiennymi. Warunkiem jaki spełniać muszą zmienne jest ich dyskretny – a często po prostu dichotomiczny – rozkład. Tabele wielodzielcze, które umożliwiają analizę liczebności odpowiadających kategoriom wyznaczanym przez więcej niż jedną zmienną, stanowią kombinację dwóch lub więcej tabel liczebności ułożonych w ten sposób, że każda komórka tabeli reprezentuje jedyną określoną kombinację konkretnych wartości tabelaryzowanych zmiennych. Tabele te konstruujemy w charakterystyczny sposób: kategorie (poziomy) jednej zmiennej ułożone są w kolumnach, a drugiej zmiennej – w rzędach tabeli. Tabelaryzować możemy jedynie zmienne dyskretne (nominalne) lub zmienne o ograniczonej liczbie sensownych wartości. Jeżeli chcielibyśmy tabelaryzować zmienną ciągłą, to należałoby ją najpierw zakodować, zamieniając wartości ciągłe na umowne rozłączne kategorie (np. niski, średni, wysoki).

Tabele liczebności wykorzystywane są w najprostszym przypadku do porównywania liczebności (jak to wynika z nazwy metody) w kilku grupach, ale dzięki analizie liczebności w komórkach tabeli liczebności można także zidentyfikować relacje, jakie zachodzą między tabelaryzowanymi zmiennymi. Chcąc ocenić takie zależności wybralibyśmy intuicyjnie metodę porównywania rozkładów liczebności w poszczególnych wierszach i kolumnach (tzw. liczebności brzegowe lub warunkowe) (*total cumulative frequencies of raws, columns*). Dokonanie takiego wstępnego porównania jest oczywiście łatwiejsze, jeżeli porównywane liczebności podawane są w formie częstości względnej (procentu).

Do oceny zależności między liczebnościami w rzędach i kolumnach tabeli wykorzystywany jest test chi^2 Pearsona (*Pearson's chi-squared test*). Statystyka tego testu jest podstawą wielu bardzo rozpowszechnionych i rutynowo stosowanych testów istotności dla zmiennych jakościowych (lub kategoryzowalnych, czyli takich, które da się zamienić na kategorie o rozkładzie dyskretnym). Wartość statystyki testu chi^2 jako miara różnic między grupami czy zależności między zmiennymi opiera się na fakcie, że istnieje możliwość szacowania liczebności oczekiwanych, to znaczy takich, jakich oczekiwalibyśmy gdyby nie istniała żadna zależność między zmiennymi. Test ten służy do badania czy rozkład przypadków pomiędzy kategoriami jednej zmiennej jest niezależny od ich rozkładu pomiędzy kategoriami drugiej zmiennej. Wartość statystyki testu chi^2 staje się rosnąco istotna w miarę powięk-

szania się odstępstw (odchyleń) od tego oczekiwanego schematu rozmieszczenia liczebności w komórkach tabeli, to znaczy w miarę jak liczebności w poszczególnych komórkach zaczynają się różnić.

Podobnie jak to miało miejsce w testach parametrycznych, wartość statystyki testu chi^2 oraz jej istotność zależy od liczebności próby, a konkretnie od liczby obserwacji i liczby komórek w tabeli: już niewielkie odchylenia od wartości oczekiwanych mogą okazać się istotne, jeżeli całkowita liczebność próby jest duża. Założenie minimalnej liczebności w poszczególnych komórkach większej od 5 jest jedynym wymaganiem i ograniczeniem tego testu: przy mniejszych liczebnościach oceny prawdopodobieństw testu są wysoce nieprecyzyjne.

W metodach wielowymiarowych, takich jak np. analiza log-liniowa, zamiast testu chi^2 Pearsona stosowany jest analogiczny test chi^2 największej wiarygodności (*chi-squared likelihood ratio test*), który testuje tę samą hipotezę co statystyka chi^2 Pearsona, jednak sposób jego obliczania oparty jest na teorii największej wiarygodności. Wartości obu tych testów są w praktyce najczęściej bardzo zbliżone.

Tabele czteropolowe

Najprostszą formą tabeli wielodzielczej jest tabela 2 na 2 (2 x 2) (*fourfold tables*), w której dwie zmienne są sklasyfikowane krzyżowo, a każda z nich może przyjmować tylko dwie wartości. Każda obserwacja należy zatem do jednej z czterech komórek tabeli. Aby zdecydować czy liczebności te są równe (kiedy nie oczekujemy związku między zmiennymi, czyli nie spodziewamy się wpływu którejkolwiek ze zmiennych na liczebność), czy też istotnie różne (kiedy zmienne nie są niezależne, czyli liczebność komórek jest różna dla różnych kategorii każdej ze zmiennych), obliczamy wartość statystyki testu chi^2, która określa jak bardzo liczebności zaobserwowane (*observed frequency*) różnią się od liczebności oczekiwanych (*expected frequency*), czyli takich, które wystąpiłyby, gdyby zmienne były całkowicie niezależne od siebie. Pod względem rachunkowym wartość chi^2 jest równa całkowitej sumie wyrażeń [(liczebność obserwowana – liczebność oczekiwana)^2/liczebność oczekiwana] liczonych dla wszystkich komórek tabeli:

$$\chi^2 = \sum \frac{\left(f_{obserwowane} - f_{oczekiwane}\right)^2}{f_{oczekiwane}}, \quad \text{dla tablicy czteropolowej } df = 1$$

Jak widać, im większa różnica między liczebnościami obserwowanymi i oczekiwanymi, tym większa wartość x^2 , i mniejsze prawdopodobieństwo, że różnice między liczebnościami są czysto przypadkowe.

W praktyce do szybkiego obliczania wartości statystyki chi^2 dla tabel 2 x 2 wygodniej jest korzystać z równania:

$$\chi^2 = \frac{n(ad-bc)^2}{efgh},$$

gdzie liczebności poszczególnych komórek odpowiadają układowi w tabeli:

grupa	*kolumna 1*	*kolumna 2*	*razem*
wiersz 1	*a*	*b*	***e***
wiersz 2	*c*	*d*	***f***
razem	***g***	***h***	***n***

Zauważmy, że tabela czteropolowa jest równoważna testom do porównywania dwóch proporcji (*zobacz niżej*).

Poprawka Yatesa

Przybliżenie statystyki chi^2 w małych tabelach typu 2 x 2 można poprawić przez pomniejszenie wartości różnicy pomiędzy oczekiwaną i zmierzoną liczebnością o 0.5 przed podniesieniem do kwadratu – jest to tzw. poprawka na ciągłość Yatesa (*Yates correction for continuity*), która sprzyja lepszej aproksymacji testu chi^2 do testu normalnego. Jest tak dlatego, że rozkład chi^2 jest rozkładem opisującym zmienne ciągłe, a wykorzystujemy go do analizy zmiennych dyskretnych, czyli nieciągłych. Czym mniejsze są liczebności w komórkach tabel liczebności, tym bardziej taka nieciągłość zmiennych się zaznacza, dlatego poprawka ta powinna być stosowana zwłaszcza wtedy, gdy liczebności w tabeli są małe – zdarza się wówczas, że niektóre liczebności oczekiwane stają się mniejsze niż 10. Proponuje się, że test chi^2 może być z powodzeniem stosowany w sytuacjach, gdy całkowita liczebność w komórkach tabeli jest nie mniejsza niż 20, a liczebności poszczególnych komórek wynoszą minimum 5 obserwacji.

Równanie opisujące wartość statystyki chi^2 z uwzględnieniem poprawki Yatesa ma postać:

$$\chi^2 = \sum \frac{\left(\left| f_{obs} - f_{oczek} \right| - \frac{1}{2} \right)^2}{f_{oczek}}, \quad d.f. = 1$$

Normalizacja rozkładu

Test normalny do porównywania proporcji oraz test chi^2 są matematycznie równocenne, ponieważ wartości statystyk obu testów są związane zależnością: $\chi^2 = z^2$. Zależność ta jest spełniona niezależnie od stosowania poprawki na ciągłość Yatesa pod warunkiem, że poprawkę taką uwzględnimy w obu porównywanych metodach obliczeń. Zauważmy, że wartości krytyczne testu chi^2 odpowiadają wartościom krytycznym obustronnego testu normalnego jedynie dla przypadku, gdy występuje 1 stopień swobody (czyli dla tabel 2 x 2). Korespondencja taka nie istnieje dla tabel wyższego rzędu (o większej liczbie pól), gdyż w takich zestawieniach mamy *de facto* do czynienia z porównaniami wielokrotnymi. Przewagą testu normalnego dla proporcji jest na pewno to, że wartości przedziału ufności można obliczyć o wiele łatwiej. Z kolei, test chi^2 jest prostszy w użyciu, a także może być on rozbudowany do porównywania więcej niż dwóch proporcji oraz do zestawień tabel wielodzielczych.

Tabele wielopolowe

Logika oraz metoda obliczeń stosowana w analizie tabel czteropolowych może być także zastosowana do tabel liczebności z większą liczbą kolumn (*c*) i wierszy (*r*):

$$\chi^2 = \sum \frac{\left(f_{obs} - f_{oczek}\right)^2}{f_{oczek}}, \quad d.f. = (r-1)\times(c-1)$$

Dla tablic wielopolowych nie istnieje opcja rachunkowa uwzględniająca poprawkę na ciągłość. Przyjmuje się, że szacowana wartość statystyki chi^2 jest poprawna, o ile nie więcej niż 20% wartości liczebności oczekiwanych jest mniejsze od 5 i żadna nie jest mniejsza od 1. Jeżeli warunek ten nie jest spełniony, praktycznym sposobem zaradzenia temu ograniczeniu może być łączenie liczebności kolumn i/lub wierszy w większe grupy.

Dla tabel wielopolowych nie dysponujemy także wariantem równania do szybkiego obliczania wartości statystyki chi^2, co najwyżej można wykorzystywać alternatywne warianty równań dla tabel 2 x *c* lub 2 x *r*. Liczebności oczekiwane należy liczyć osobno dla każdej komórki według wzoru:

$$f_{oczek} = \frac{(suma\ kolumny)\times(suma\ wiersza)}{suma\ cakowita}$$

Chociaż w celu wstępnego porównania liczebności w komórkach tabeli przedstawia się je często w formie częstości względnych (procentu), częstości takich nie należy stosować przy liczeniu wartości statystyki chi^2, wtedy zawsze stosujemy liczebności bezwzględne; odnosi się to zresztą do wszystkich rodzajów tabel wielodzielczych, także tabel czteropolowych.

Test chi^2 dla tabeli typu 2 x *c*, czyli takiej, która zawiera dwa wiersze i *c* kolumn, jest niczym innym jak testem *c* proporcji, z których każda reprezentuje jedną kolumnę tabeli. W takim przypadku do obliczania wartości statystyki chi^2 możemy posłużyć się wzorem:

$$\chi^2 = \frac{N^2\left[\sum\left(r^2/n\right) - R^2/N\right]}{R(N-R)}, \quad d.f. = c-1$$

gdzie *n* jest sumaryczną liczebnością dla danej kolumny, zaś *r* liczebnością górnego rzędu z tej kolumny; podobnie, *N* jest całkowitą liczebnością tabeli, zaś *R* sumaryczną liczebnością górnego rzędu w tabeli. Ponieważ w analizie tabel liczebności wiersze można traktować wymiennie z kolumnami bez wpływu na wartość szacowanej statystyki chi^2, podany powyżej wzór można stosować jednakowo dobrze także do tabel typu *r* x 2.

Test dokładny Fishera

Jeżeli liczebności są niewielkie, to aproksymacja testu chi^2 do testu normalnego nie jest wiarygodna. Obliczana wartość statystyki chi^2 nie jest wtedy dobrym wymiernikiem zależności między zmiennymi. W sytuacjach gdy:

- całkowita liczebność jest mniejsza niż 20, lub
- całkowita liczebność wynosi między 20 a 40, i najmniejsza z liczebności jest niższa niż 5, zamiast testu chi^2 stosujemy tzw. dokładny test Fishera (*Fisher's exact test*).

Wartość tzw. dokładnego prawdopodobieństwa testu Fishera dla tablicy czteropolowej obliczamy według równania:

$$P_{(2\times2)} = \frac{e!\,f!\,g!\,h!}{n!\,a!\,b!\,c!\,d!}$$

gdzie symbole a, b, c, d, e, f, g, h są liczebnościami poszczególnych komórek tabeli 2 x 2 zgodnie z oznaczeniem podanym poniżej.

grupa	*kolumna 1*	*kolumna 2*	*razem*
wiersz 1	*a*	*b*	***e***
wiersz 2	*c*	*d*	***f***
razem	***g***	***h***	***n***

Pamiętajmy, że test ten jest dostępny jedynie w tabelach 2 x 2 i nie ma odpowiedników w tabelach wielopolowych. Opiera się on na obliczaniu, jakie są prawdopodobieństwa wystąpienia liczebności obserwowanych oraz prawdopodobieństwa wszystkich bardziej skrajnych liczebności od tych, które występują w tabeli, przy zachowaniu tych samych liczebności sumarycznych dla wierszy i kolumn (zobacz *„Część II – Uzupełnienia, przykłady i zadania"*).

Istotność różnic badaną tym testem można w praktyce ocenić dwojako:

(*i*) jest to albo suma prawdopodobieństwa wystąpienia liczebności takich jak te obserwowane w tabeli plus prawdopodobieństw wystąpienia liczebności bardziej skrajnych (czyli mniej prawdopodobnych tabel):

istotność (*metoda I*) = P (*liczebności obserwowane w tabeli*) + P (*liczebności mniej prawdopodobne*)

albo

(*ii*) podwojona suma prawdopodobieństwa wystąpienia liczebności takich jak te obserwowane w tabeli plus prawdopodobieństw wystąpienia liczebności bardziej skrajnych (czyli mniej prawdopodobnych tabel) przy zachowaniu tego samego kierunku zmian, co obserwowany w tabeli:

istotność (*metoda II*) = 2 x [P (*liczebności obserwowane w tabeli*) + P (*liczebności bardziej skrajnego rozkładu przy zachowaniu tego samego kierunku zmian*)]

Testy sparowane (*chi-squared test for paired case*)

Porównywanie dwóch proporcji

W niektórych badaniach jesteśmy zainteresowani porównaniami proporcji obliczanymi na podstawie obserwacji sparowanych. Mogą to być proporcje zmierzone dwukrotnie u tych samych osobników, lub badania kontrolno-kliniczne, gdy osoby włączane do badań dobieramy w ściśle określony sparowany sposób, na przykład pod względem wieku i/lub płci. Pod względem rachunkowym metody te opierają się na logice testu chi^2 dla tabel czteropolowych (dla liczebności) lub testu normalnego (dla proporcji), ale charakteryzują się one własną odmienną specyfiką zestawienia danych w tabelach liczebności, która jest różna od typowych tabel 2 x 2 (zobacz *„Część II – Uzupełnienia, przykłady i zadania"*).

Test chi² McNemara – test niezgodnych par

Test chi² McNemara (*McNemar's chi-squared test, discordant pairs test*) należy stosować, jeśli liczebności w tabeli 2 x 2 reprezentują próbki zależne. Typowymi zastosowaniami tego testu jest porównywanie liczebności w układzie pomiarowym typu przed i po (np. operacji, kuracji, zadziałaniu czynnika, itp.). Statystyka chi² testu McNemara, uwzględniająca liczbę par niezgodnych *b* i *c*, testuje hipotezę, że liczebności w komórkach *b* i *c* (prawa górna i lewa dolna) są identyczne:

	+	–	*razem*
+	***a***	*b*	*e*
–	*c*	***d***	*f*
razem	*g*	*h*	*n*

$$\chi^2_{par} = \frac{(|b-c|-1)^2}{b+c}, \quad d.f. = 1$$

Test McNemara, lub jego aproksymację do rozkładu normalnego, można z powodzeniem wykorzystywać, o ile liczba par niezgodnych wynosi przynajmniej 10. W przypadku mniejszej liczby par należy zastosować metodę szacowania dokładnych prawdopodobieństw dla zmiennych binarnych (rozkład dwumianowy) (zobacz przykład 83 w *„Części II – Uzupełnienia, przykłady i zadania"*).

Różnicę proporcji sparowanych, która jest liczbowo równa różnicy proporcji wyników dodatnich (czyli proporcji zgodnych wyników w stosunku do wszystkich obserwowanych wyników), błąd standardowy takiej różnicy oraz odpowiedni przedział ufności, można policzyć według równań:

$$różnica = \frac{b-c}{n} \qquad SE = \frac{\sqrt{b+c}}{n}$$

$$CI = \frac{(b-c)}{n} \pm z' \frac{\sqrt{(b+c)}}{n}$$

Tabele zbiorcze 2 x 2

W przypadku analiz wielu różnych tabel czteropolowych ich łączenie może prowadzić do przekłamania wyników testu chi² we wszystkich przypadkach, gdzie mamy do czynienia z występowaniem zmiennych uwikłanych (towarzyszących, stanowiących dodatkowy czynnik), które mogą wpływać na liczebności komórek tabeli. Taką zmienną jest bardzo często płeć lub wiek. Łączenie danych może istotnie maskować zasadniczy efekt, jeżeli nasilenie tego efektu zależy istotnie od wartości zmiennej towarzyszącej. Optymalnie byłoby przeprowadzić analizę w każdej z mniejszych podgrup osobno. Jeżeli jednak chcemy szacować efekt w całej grupie, a nie osobno w każdej z podgrup, należy wtedy stosować metody standaryzacji danych opisane dokładnie w dalszej części tego opracowania (*„Metody wykorzystywane w badaniach populacyjnych"*).

Test chi² Mantela-Haenszela

Niekiedy zachodzi potrzeba policzenia wartości sumarycznej statystyki chi² dla układu zawierającego kilka oddzielnych podgrup danych (np. różne płcie, różne grupy wiekowe, różne stadia zaawansowania choroby, itp.). Metodą, która na to pozwala, zachowując jednocześnie informacje na temat wpływu towarzyszących zmiennych na rozkład liczebności w takiej zbiorczej tabeli składającej się z kilku niezależnych tabel czteropolowych, jest test chi² Mantela-Haenszela (*Mantel-Haenszel chi-squared test*). Test ten jest szczególnie przydatny w sytuacjach, gdy wpływ badanych zmiennych jest widoczny, ale nieistotny, na przykład ze względu na małe liczebności w podgrupach. Aby policzyć wartość statystyki chi² tego testu musimy znać:

- liczebność obserwowaną, a,
- liczebność oczekiwaną dla a, $f_{oczek(a)} = eg / n$,
- zmienność liczebności a, $V_a = efgh/[n^2(n-1)]$

zgodnie z oznaczeniami komórek tabeli:

grupa	kolumna 1	kolumna 2	razem
wiersz 1	*a*	*b*	*e*
wiersz 2	*c*	*d*	*f*
razem	*g*	*h*	***n***

Wartość statystyki chi² z uwzględnieniem poprawki na ciągłość obliczamy następująco:

$$\chi^2_{MH} = \frac{(|\sum a - \sum f_{oczek(a)}| - 0.5)^2}{\sum V_a}, \; d.f. = 1$$

Zauważmy, że statystykę dla tego testu oceniamy dla jedynie 1 stopnia swobody, niezależnie od tego, ile tabel 2 x 2 zawiera nasza tabela zbiorcza.

Może się wydawać dziwne, że wartość statystyki chi² jest liczona jedynie w oparciu o liczebność w komórce a, bez uwzględniania liczebności innych komórek tabeli. Zauważmy jednak, że w sytuacji gdy znamy wartość a, możemy policzyć także liczebności dla innych komórek na podstawie liczebności brzegowych e, f g oraz h. Gdybyśmy zastosowali test Mantela-Haenszela do standardowej tabeli 2 x 2, szacowana wartość statystyki chi² byłaby bliska (ale nie identyczna) wartości standardowego testu chi²: dokładnie – byłaby ona $(n-1)/n$ razy mniejsza od wartości standardowej chi². Łatwo zauważyć, że różnica ta jest zaniedbywalnie mała dla liczebności całkowitych powyżej 20 (< 5%), czyli takich jakie stanowią graniczną liczebność dla stosowania testu chi².

Ograniczenia

Test Mantela-Haenszela jest aproksymacją, a poprawność jego stosowania w konkretnych przypadkach określa tzw. „reguła 5". Aby ją zastosować musimy znać sumy następujących wyrażeń dla wszystkich komórek tabeli:

(*i*) wartość minimalna (e, g), oraz

(*ii*) wartość maksymalna (0, $g - f$), która wynosi 0 gdy g jest mniejsze lub równe f lub $g - f$, gdy g jest większe.

Obie sumy muszą różnić się od sumy liczebności oczekiwanych przynajmniej o 5, aby można było zastosować test Mantela-Haenszela.

Test chi² dla trendu

Test chi² dla tabel postaci 2 x *c* jest ogólnym testem wykorzystywanym do badania różnic między *c* proporcjami, z których każda reprezentuje jedną kolumnę tabeli. Jeżeli kolumny takie układają się w określonym porządku rosnącym (lub malejącym), to lepszym rozwiązaniem jest zastosowanie testu chi² dla trendu (*chi-squared test for trend*) niż zwykłego testu porównania proporcji:

$$\chi^2_{trend} = \frac{\left(|A| - 0.5\right)^2}{B}, \quad d.f. = 1,$$

(równanie uwzględnia poprawkę na ciągłość, którą można pominąć, gdy punktacja komórki nie jest tożsama z numerem porządkowym kolumny)

$$\text{gdzie} \quad A = \sum(rx) - \frac{R}{N}\sum(nx) \quad \text{i} \quad B = \frac{R(N-R)}{N^2(N-1)}\left[N\sum(nx^2) - (\sum nx)^2\right]$$

W równaniach tych: *rx* oznacza iloczyn liczebności górnego wiersza każdej kolumny i punktacji komórki wyznaczającej jej miejsce w porządku rosnącym (lub malejącym) badanego trendu, *nx* i nx^2 to iloczyny punktacji (lub jej kwadratu) komórki dla badanego trendu i sumarycznej liczebności odpowiedniej kolumny, *N* i *R* oznaczają odpowiednio liczebność całkowitą tabeli oraz liczebność sumaryczną dla górnego wiersza.

Jeżeli dane obejmują kilka oddzielnych podgrup i nie można ich połączyć ze względu na obecność zmiennych uwikłanych, test trendu należy przeprowadzać osobno dla każdego podzbioru danych.

Można jednak w takim przypadku – podobnie jak to zrobiliśmy dla tabel zbiorczych 2 x 2 – zastosować całościowy test dla trendu, który uwzględnia zarówno indywidualne trendy w podgrupach, jak i (towarzyszące) zmienne uwikłane. Niezależnie od liczby podgrup test taki ma tylko 1 stopień swobody:

$$\chi^2_{całościowy} = \frac{\left(\sum|A| - 0.5\right)^2}{\sum B}$$

Rzadziej stosowane statystyki oparte na teście chi², takie jak współczynnik *F* czy współczynnik zgodności, omówiono w Rozdziale „*Metody nieparametryczne*".

Podsumowanie

- Tabele liczebności służą do porównywania liczebności grup oraz do identyfikowania relacji między zmiennymi jakościowymi (lub takimi, które da się zamienić na kategorie o rozkładzie dyskretnym).

- W analizie tabel liczebności stosujemy test chi^2 Pearsona, który służy najczęściej do badania, czy rozkład przypadków należących do różnych kategorii jednej zmiennej jest niezależny od rozkładu dla kategorii drugiej zmiennej. Wartość statystyki tego testu jest miarą różnic między grupami lub zależności między zmiennymi.
- Statystyka testu chi^2 Pearsona, stosowana w analizie danych o rozkładach nieciągłych, jest oparta na ciągłym rozkładzie chi^2. Przybliżenie statystyki ciągłego rozkładu chi^2 dla zmiennych dyskretnych (nieciągłych) można poprawić stosując tzw. poprawkę na ciągłość Yatesa, która sprzyja lepszej aproksymacji testu chi^2 do testu normalnego. Poprawkę tę stosujemy wyłącznie w małych tabelach 2 x 2.
- Wartość statystyki testu chi^2 oraz jej istotność zależy od liczby obserwacji i liczby komórek w tabeli. Założenie minimalnej liczebności w poszczególnych komórkach większej od 5 oraz całkowitej liczebności nie mniejszej niż 20 jest podstawowym wymaganiem i ograniczeniem tego testu.
- W sytuacjach gdy założenie to nie jest spełnione, powinniśmy zamiast testu chi^2 stosować dokładny test Fishera.
- Test chi^2 dla tabel 2 x 2 jest matematycznie równocenny testowi normalnemu do porównywania proporcji. Przewagą testu normalnego dla proporcji jest łatwość oszacowania przedziału ufności, test chi^2 z kolei może być rozbudowany do porównywania więcej niż dwóch proporcji oraz do zestawień tabel wielodzielczych.
- Tabele wielodzielcze, które umożliwiają analizę liczebności odpowiadających kategoriom wyznaczanym przez więcej niż jedną zmienną, stanowią kombinację dwóch lub więcej tabel liczebności typu 2 x 2.

Rozdział 11

Metody wielowymiarowe

Z uwagi na niezwykłą złożoność i wielowymiarowość otaczającej nas rzeczywistości sytuacje, w których pojedyncza zmienna jest w stanie opisać i wyjaśnić dane zjawisko (tak jak ma to miejsce np. przy stosowaniu testu *t*) należą do rzadkości. W przypadku kiedy mamy do czynienia jednocześnie z więcej niż jedną zmienną zależną lub niezależną, logika i istota obliczeń nie zmienia się w stosunku do analiz jednowymiarowych, chociaż obliczenia stają się coraz bardziej złożone. Do analizy statystycznej takich przypadków stosujemy różne metody wielowymiarowe (*multiparametric analyses*), które pozwalają nam badać jak na zmienną lub zmienne zależne wpływa wiele zmiennych wyjaśniających, czyli w jaki sposób liczne zmienne zależne zmieniają się razem. Do metod tych należą na przykład analiza wariancji, regresja wielokrotna, regresja logistyczna czy modele log-liniowe.

W przypadkach kiedy mamy kilka/wiele zmiennych zależnych nasza hipoteza zakłada, że wszystkie zmienne ulegają wpływowi różnicy pomiędzy grupami (kategoriami zmiennej lub zmiennych niezależnych). Do testowania takich hipotez możemy na przykład wykorzystać różne rodzaje wielowymiarowej analizy wariancji (*multiple multiparametric analysis of variance*) (MANOVA). Jeśli nasz globalny test wielowymiarowy wykryje istotność między porównywanymi grupami, wówczas wnioskujemy, że odpowiedni efekt (wpływ zmiennej grupującej) jest istotny. Nasze dalsze pytania mogą dotyczyć tego, która ze zmiennych zależnych – i na ile – jest istotna w globalnym kształtowaniu tego efektu.

W praktyce po stwierdzeniu statystycznej istotności testu wielowymiarowego (globalnego efektu głównego) możemy przeprowadzić jednowymiarowe testy *F* dla każdej ze zmiennych, aby dokonać interpretacji odpowiedniego efektu, czyli zidentyfikować zmienną zależną, która wnosi wkład w istotność ogólnego efektu. Należy pamiętać jednak, że taka jednowymiarowa analiza nie pozwala nam na ocenę kompleksową – na określenie czy jedna grupa różni się od drugiej zespołem cech. W takich przypadkach stosujemy właśnie analizy wielowymiarowe – przeprowadzane w wielu wymiarach, to jest z uwzględnieniem wielu zmiennych równocześnie. Taka *„wielowymiarowość"* jest powodem, że metodologicznie i obliczeniowo metody te są złożone i analizy tego typu przeprowadza się zawsze z wykorzystaniem komputerowych pakietów statystycznych.

Analiza dyskryminacji

Analiza dyskryminacyjna (*discriminant analysis*) jest bardzo przydatnym narzędziem służącym do:

- wykrywania tych zmiennych, które pozwalają badaczowi najlepiej dyskryminować (rozdzielać) różne (naturalne wyłaniające się) grupy, oraz
- do klasyfikacji poszczególnych obserwacji (przypadków) do różnych grup z większą trafnością niż przez czysty przypadek.

Należy ona do tzw. analiz wielowymiarowych, to znaczy takich, które uwzględniają wpływ grupy zmiennych (a nie jednej zmiennej) na inną zmienną w tym samym czasie. Na przykład, wiemy, że pacjenci z chorobą wieńcową charakteryzują się zmienionymi wartościami licznych parametrów biochemicznych, metabolicznych, koagulologicznych, reologicznych, itd. Badając zmienność każdego z tych parametrów możemy wykazać (lub nie) istotne różnice ich wartości w obu grupach, tzn. u ludzi zdrowych i osób z chorobą wieńcową. Niektóre z nich odróżniają badane grupy bardzo wyraźnie, tzn. mają wysoce istotnie różne wartości średnie, inne nie różnią się na tyle istotnie, abyśmy postrzegali te różnice jako regularną prawidłowość spotykaną zawsze przy porównaniach osób zdrowych i osób chorych. Wypowiadając się na przykład na temat tego, czy porównywane grupy różnią się pewnym zespołem parametrów biochemicznych czy koagulologicznych możemy jedynie – na podstawie indywidualnego badania istotności różnic dla każdego z parametrów – orzec, że niektóre parametry są istotnie różne między grupami, a inne nie. Jest tak dlatego, iż analizę przeprowadzamy jakby w jednym wymiarze – niezależnie dla każdego parametru – nie wnikając w możliwe interakcje między zmiennymi zależnymi.

Analiza funkcji dyskryminacyjnej jako narzędzie do dyskryminacji grup

Analiza funkcji dyskryminacyjnej jest stosowana do rozstrzygania, czy dany „zestaw" zmiennych – i które zmienne w takim zestawie – dyskryminują dwie lub więcej naturalnie wyłaniające się grupy. Co więcej, analizę dyskryminacyjną można wykorzystać do rozstrzygania, która zmienna lub zmienne przyczyniają się w największym stopniu do rozseparowania grup, a także które z nich mogą być najlepszymi predyktorami przynależności określonych obserwacji (przypadków) do danej grupy.

Pod względem rachunkowym, analiza funkcji dyskryminacyjnej jest bardziej ogólną formą metody analizy wariancji (ANOVA). Główną ideą analizy funkcji dyskryminacyjnej jest rozstrzyganie, czy grupy różnią się ze względu na średnią pewnej zmiennej, a następnie wykorzystanie tej zmiennej do przewidywania przynależności nowych przypadków do określonej grupy. Jeżeli określona zmienna przyjmuje zasadniczo różne wartości dla przypadków należących do różnych grup, to już ta jedna zmienna może być użytecznym predyktorem przynależności do określonej grupy, ponieważ jej wartość dobrze dyskryminuje porównywane grupy. Najczęściej jednak pojedyncze zmienne nie dają tak doskonałej dyskryminacji i musimy wykorzystać zespół cech (parametrów, zmiennych) w celu właściwej i poprawnej klasyfikacji.

Jeżeli mamy do czynienia z pojedynczą zmienną, to stosujemy test F do weryfikacji czy zmienna ta jest różna w różnych grupach, tzn. czy zmienna ta dyskryminuje grupy. Wartość statystyki testu F zależy oczywiście od wielkości wariancji międzygrupowej: jeśli jest ona istotnie większa od wariancji wewnątrzgrupowej, to muszą występować istotne różnice

między średnimi w badanych grupach. Ponieważ w metodzie analizy dyskryminacyjnej mamy do czynienia z wieloma zmiennymi, zamiast wartości jednowymiarowej statystyki *F* obliczamy wartość wielowymiarową *F* (tzw. lambda Wilksa), opartą na porównaniu macierzy wariancji błędu (zmienności wewnątrzgrupowej) i macierzy efektu (zmienności międzygrupowej).

Jednym z dwóch najpowszechniejszych zastosowań metody analizy funkcji dyskryminacyjnej jest wyodrębnianie spośród wielu zmiennych (cech, parametrów) tych, które najlepiej dyskryminują grupy. Znajomość takich zmiennych pozwala nam dobierać je w celu sprawdzania przynależności poszczególnych obserwacji do określonej grupy. Mówiąc o zmiennych znajdujących się w modelu mamy na myśli te, które najbardziej przyczyniają się do poprawnej separacji grup, podczas gdy pozostałe zmienne znajdujące się poza modelem, nie wnoszą żadnej poprawy dyskryminacji porównywanych grup.

Taki najwłaściwszy model analizy dyskryminacyjnej budowany jest etapowy: do modelu włączane są stopniowo kolejne zmienne w oparciu o ocenę czy dana zmienna poprawia (lub nie pogarsza) dyskryminacji grup (jest to tak zwany model analizy dyskryminacyjnej krokowej postępującej). Alternatywnie, do modelu włączane są wszystkie zmienne, a następnie, zmienne, które najmniej wnoszą do przewidywania przynależności do grupy, są stopniowo eliminowane z modelu. Ostatecznym wynikiem analizy jest włączenie wszystkich „istotnych" zmiennych (czyli takich, które najbardziej przyczyniają się do dyskryminacji grup), oraz pozbycie się zmiennych zbędnych. „*Selekcja*" taka dokonuje się w oparciu o wartość statystyki testu *F* dla każdej zmiennej, która wskazuje na istotność statystyczną zmiennej w dyskryminowaniu grup, to znaczy mówi, jaki jest indywidualny przyczynek zmiennej w przewidywaniu przynależności danej obserwacji do grupy. Takie ważenie „znaczenia" każdej zmiennej przy dużej liczbie zmiennych i przypadków jest dość zawiłą procedurą pod względem rachunkowym i nie byłoby oczywiście możliwe bez wykorzystania komputerowych pakietów statystycznych. Musimy sobie uświadamiać, że tak liczona istotność nie jest oczywiście odzwierciedleniem rzeczywistej wielkości błędu I rodzaju, czyli nie mówi nam dokładnie jakie jest prawdopodobieństwo błędnego odrzucenia hipotezy zerowej, mówiącej, że nie ma żadnego zróżnicowania między grupami. Dzieje się tak dlatego, że algorytm działania takich komputerowych procedur statystycznych polega na „przebieraniu" wśród wielu zmiennych i poszukiwaniu modelu, w którym dyskryminacja między grupami będzie najlepsza.

Ogólnie, analiza funkcji dyskryminacyjnej jest metodą analogiczną do metody regresji wielokrotnej, a takie podobieństwo widać wyraźnie, kiedy porównujemy dwie grupy. W przypadku takim, równanie regresji wielokrotnej odpowiada dokładnie równaniu funkcji dyskryminacyjnej i ma postać:

$$y = a + b_1x_1 + b_2x_2 + ... + b_nx_n$$

Interpretując wyniki takiego równania pamiętamy, że zmienne, które mają największe (standaryzowane) współczynniki regresji (*standardized regression coefficients*), przyczyniają się najbardziej do predykcji zmiennej zależnej na podstawie zmiennych niezależnych. Jeżeli naszą zmienną zależną zdefiniujemy sobie jako przynależność obserwacji do którejś grupy, to możemy powiedzieć, że zmienne o największych (standaryzowanych) współczynnikach regresji najbardziej przyczyniają się do predykcji przynależności do grupy. Współczynniki funkcji dyskryminacyjnej określają zatem cząstkowy (indywidualny) wkład każdej zmiennej do funkcji dyskryminacyjnej. Jeżeli grup jest więcej niż dwie, to rośnie liczba funkcji

dyskryminacyjnych (tzw. pierwiastków). Funkcji tych może być maksymalnie tyle, ile wynosi liczba grup minus jeden lub liczba zmiennych włączonych do analizy minus jeden, w zależności od tego, która z tych liczb jest mniejsza. Procedury obliczania takich funkcji dyskryminacyjnych charakteryzują się tym, że każda z funkcji jest niezależna od innych – udział tych funkcji w dyskryminowaniu grup nie pokrywa się. Mówimy, że funkcje takie są ortogonalne (niezależne). Przy estymacji poszczególnych funkcji dyskryminacyjnych kombinacja zmiennych przyczyniających się do jak najlepszej dyskryminacji grup jest zoptymalizowana w taki sposób, że pierwsza określona funkcja opisuje najbardziej ogólne rozróżnienie między wszystkimi grupami, druga bardziej szczegółowe (z naciskiem na separację pewnych grup od innych), następna jeszcze bardziej szczegółowe, itd. Taka optymalizacja dokonywana jest w oparciu o wyniki korelacji kanonicznej. O tym jak dana funkcja dyskryminuje określone grupy możemy dowiedzieć się porównując średnie dla tych funkcji w poszczególnych grupach. Możemy także zobrazować graficznie, w jaki sposób zestawienie dwóch określonych funkcji dyskryminuje grupy w układzie współrzędnych opisującym zależność dwóch funkcji dyskryminacyjnych. Aby ocenić, które zmienne mają największy udział w zdefiniowaniu określonej funkcji dyskryminacyjnej, oblicza się tzw. współczynniki struktury czynnikowej, które oznaczają korelacje między zmiennymi w modelu a funkcjami dyskryminacyjnymi (są to jakby ładunki czynnikowe poszczególnych zmiennych w funkcji dyskryminacyjnej, analogicznie jak w metodzie analizy czynnikowej).

Przy dużej liczbie grup liczba zbudowanych funkcji dyskryminacyjnych może być duża. Może się okazać, że nie każda jest istotna, czyli przyczynia się zauważalnie do polepszenia dyskryminacji między grupami. Nieistotne funkcje dyskryminacyjne to takie, które nie polepszają dyskryminacji, czyli możemy je zignorować. Pamiętamy, że jedynie początkowe funkcje dyskryminacyjne (pierwsza, druga, itd.) mają charakter ogólny i dotyczą wszystkich porównywanych grup. Dalsze są bardziej szczegółowe i dlatego ich wpływ na separację wszystkich grup między sobą staje się coraz mniej istotny, ich moc dyskryminacyjna maleje.

Analiza funkcji dyskryminacyjnej jako narzędzie do klasyfikacji przypadków

Predykcja przynależności obserwacji do poszczególnych grup jest jednym z głównych celów zastosowania analizy dyskryminacyjnej. Dokonujemy jej w oparciu o obliczone funkcje dyskryminacyjne (predykcja *post hoc*) lub deklarujemy taką przynależność wstępnie przy konstruowaniu modelu (predykcja *a priori*). Ta druga predykcja służy zbudowaniu pewnego wzorca (matrycy), który wykorzystujemy następnie do klasyfikacji nowych obserwacji. W sytuacji, gdy jakiś zbiór danych służy nam do estymacji funkcji dyskryminacyjnych, które najlepiej dyskryminują grupy, a następnie wykorzystujemy te same dane do oceny, na ile trafna jest nasza predykcja, to uzyskamy zawsze lepszą klasyfikację niż wtedy, gdy przewidujemy przynależność przypadków, które nie były użyte do estymacji funkcji dyskryminacyjnej. Czyli nasza predykcja *post hoc* będzie zawsze lepsza niż predykcja *a priori*, ponieważ o wiele łatwiej wypowiedzieć się o tym co się zdarzyło, niż przewidzieć to co się stanie. W praktyce ocena poprawnej klasyfikacji dotyczy z reguły przyszłych obserwacji, nie zaś obserwacji z tego samego zbioru, na podstawie którego oszacowano funkcje dyskryminacyjne.

Aby ułatwić rozstrzyganie, do której grupy najprawdopodobniej należy dany przypadek, metoda analizy dyskryminacyjnej umożliwia obliczanie tzw. funkcji klasyfikacyjnych (*classification function*), których jest tyle, ile porównywanych grup (czyli każda grupa ma przyporządkowaną sobie funkcję klasyfikacyjną, która ją opisuje). Funkcja taka posiada ogólną postać:

$$S_i = c_i + w_{i1}x_1 + w_{i2}x_2 + ... + w_{im}x_m$$

gdzie i określa numer danej grup, indeksy 1, 2,..., m określają m zmiennych, c_i jest stałą dla grupy i, w_{ij} jest współczynnikiem dla zmiennej j w grupie i, x_j jest wartością obserwowaną dla danego przypadku zmiennej j, natomiast S_i jest wypadkową wartością funkcji klasyfikacyjnej. Funkcje klasyfikacyjne wykorzystujemy do obliczania wartości klasyfikacyjnych dla poszczególnych (także nowych) obserwacji (przypadków). Na podstawie takich wartości możemy sklasyfikować określony przypadek jako należący do grupy, dla której ma on największą wartość klasyfikacyjną (w sytuacji, gdy prawdopodobieństwa klasyfikacji *a priori* nie różnią się poważnie).

Możemy także obliczyć prawdopodobieństwo tego, iż dany przypadek rzeczywiście znajdzie się w grupie, do której został on zaszeregowany. Takie prawdopodobieństwa obliczamy oczywiście *post hoc*, po wykonaniu analizy i obliczeniu funkcji dyskryminacyjnych, są one prawdopodobieństwami *a posteriori*. Do ich oszacowania służą nam tzw. odległości Mahalanobisa (*Mahalanobis' distance*). Są to miary oddalenia od siebie zbiorowości punktów reprezentujących poszczególne grupy w przestrzeni wielowymiarowej. Gdybyśmy mieli do czynienia z pojedynczą zmienną, to miarą podobieństwa (lub niepodobieństwa) dwóch dowolnych punktów byłaby liniowa odległość między nimi w dwuwymiarowym układzie współrzędnych *xy*. W takiej sytuacji nasza odległość Mahalanobisa odpowiadałaby odległości euklidesowej. Skoro jednak zmiennych jest wiele, to wymiarowość takiego układu rozrasta się – jest on tylu-wymiarowy ile zmiennych obejmuje nasz model. Odległości Mahalanobisa oddają poza tym w sposób bardziej wiarygodny oddalenie od siebie zbiorowości punktów dla różnych grup w przypadku, gdy zmienne nie są całkowicie niezależne (ortogonalne, *orthogonal variables*), czyli miara ta zawiera w sobie także informację o skorelowaniu zmiennych. Dla każdej zbiorowości punktów charakteryzujących określoną grupę, można wyznaczyć matematyczny środek – będzie to punkt reprezentujący średnie ze wszystkich zmiennych w przestrzeni wielowymiarowej. Ten punkt nazywa się centroidą grupy. Znając położenie punktu centralnego (centroidy, *centroid*), jak również położenie dowolnego punktu dla dowolnego przypadku, możemy zaklasyfikować dany przypadek do grupy na podstawie jego odległości od centroidy. Logicznie, należałby on do tej grupy, której jest najbliższy, to znaczy do tej, do której odległość Mahalanobisa jest najmniejsza.

Możemy więc uznać, że prawdopodobieństwo, iż przypadek zostanie włączony do danej grupy jest zasadniczo proporcjonalne do odległości Mahalanobisa od punktu centralnego grupy (nie jest to dokładna proporcjonalność z uwagi na wielowymiarowy przedział rozkładów normalnych wielu zmiennych wokół każdej centroidy). Położenie każdego przypadku w takiej wielowymiarowej przestrzeni obliczamy na podstawie wcześniejszej analizy i oszacowaniu wartości, jakie zmienne należące do modelu przyjmują dla danego przypadku. Takie prawdopodobieństwa przynależności przypadku do grupy nazywamy prawdopodobieństwami *a posteriori*.

Dodatkowym czynnikiem wpływającym na jakość predykcji podczas klasyfikacji przypadków jest liczba obserwacji w grupach. Jeżeli w jednej z grup jest więcej obserwacji niż w jakiejś innej; to możemy założyć, że prawdopodobieństwo *a priori*, iż przypadek należy do tej grupy jest większe (ponieważ częściej występują w przyrodzie przedstawiciele liczniejszej grupy). W takiej sytuacji nierówna liczebność grup będzie odzwierciedleniem rzeczywistego rozkładu w populacji. Niekiedy może być ona jednak jedynie wynikiem doboru losowego próby (czynnikiem losowym). W pierwszym przypadku nasze deklarowane prawdopodobieństwa *a priori* powinny być proporcjonalne do rozmiarów grup w naszej próbie, w drugim zaś – powinny być jednakowe dla każdej grupy. Właściwe dobranie i specyfikacja prawdopodobieństw *a priori* jest bardzo istotne przy tworzeniu modelu analizy funkcji dyskryminacyjnej, gdyż może poważnie wpłynąć na trafność predykcji.

Ograniczenia i wymagania metody

Metoda analizy funkcji dyskryminacyjnej zakłada, że dane (dla poszczególnych zmiennych) reprezentują próbę z wielowymiarowego rozkładu normalnego. I tutaj, naruszanie tego założenia nie jest zazwyczaj krytyczne, ponieważ odpowiednie wypadkowe testy istotności są dość odporne na odstępstwa od normalności rozkładu. Podobnie jest z założeniem o homogeniczności wariancji: nieznaczne odchylenia od tego założenia nie są tak *„zgubne"*.

Istotnym zagrożeniem dla trafności testów istotności w tej metodzie jest przypadek skorelowania średnich w grupach z wariancjami, szczególnie przy dużych wartościach średnich niektórych zmiennych. Sytuacja taka pojawia się często wtedy, gdy mamy do czynienia z obserwacjami bardzo odstającymi od reszty w grupie. Jeżeli do analizy włączamy zbyt wiele zmiennych silnie zależnych od siebie (a więc silnie ze sobą skorelowanych), to macierz naszych danych cząstkowych może się okazać źle uwarunkowana, to znaczy nie będą możliwe do przeprowadzenia zabiegi matematyczne (takie jak na przykład odwracanie macierzy) niezbędne do obliczeń modelu analizy wariancji. Mówimy o takich zmiennych, że są one wysoce zbędne (redundantne, *redundant variables*), to znaczy ich obecność nie wnosi nowych porcji informacji do modelu. Jest to oczywiste, gdyż takie bardzo silne skorelowanie (zależność) zmiennych przypomina sytuację występowania *„replik"* (duplikatów) tej samej zmiennej.

Pojęcie o tym, które ze zmiennych są wysoce zbędne, daje nam wartość współczynnika **tolerancji** (*tolerance*), obliczaną jako 1 minus R^2 danej zmiennej przy włączeniu do bieżącego modelu wszystkich innych zmiennych. Jeżeli wartość R^2 jest wysoka (i odpowiednio tolerancja niska), oznacza to, że dana zmienna jest silnie skorelowana z innymi zmiennymi (czyli część wariancji swoista dla danej zmiennej jest niewielka, gdyż tłumaczą ją inne zmienne w modelu). Ogólnie, czym bardziej zbędna (redundantna) jest zmienna, tym bliższa zeru jest jej wartość tolerancji. Możemy z tego wnosić, że najbardziej użyteczne będą dla nas te zmienne, których tolerancja jest wysoka (a zbędność niska).

Podsumowanie

- Przy interpretacji wielokrotnych funkcji dyskryminacyjnych, z którymi mamy do czynienia wtedy, gdy analiza dotyczy więcej niż dwóch grup i więcej niż jednej zmiennej, powinniśmy przetestować istotność statystyczną różnych funkcji i dalszej analizie uwzględniać tylko funkcje istotne.
- Standaryzowane współczynniki kierunkowe każdej zmiennej w każdej istotnej funkcji modelu mówią nam o wkładzie poszczególnych zmiennych w dyskryminację grup: im większa wartość standaryzowanego współczynnika, tym większy indywidualny wkład odpowiedniej zmiennej w dyskryminacji określonej przez daną funkcję dyskryminacyjną.
- Porównanie średnich dla istotnych funkcji dyskryminacyjnych pomaga nam się zorientować, które grupy są najlepiej rozróżniane przez dane funkcje.
- Miarą separacji (dyskryminacji) między grupami mogą być odległości Mahalanobisa, czyli oddalenie w przestrzeni wielowymiarowej punktów centralnych (centroid) poszczególnych grup.
- W celu oszacowania szansy, że dany przypadek należy do konkretnej grupy obliczamy prawdopodobieństwo *a posteriori*, oparte na naszej wiedzy o wartościach dla innych przypadków.
- Funkcje klasyfikacyjne wykorzystujemy do obliczania wartości klasyfikacyjnych dla poszczególnych obserwacji (przypadków). Na ich podstawie możemy sklasyfikować określony przypadek jako należący do grupy, dla której ma on największą wartość klasyfikacyjną.
- Jedynie klasyfikacja nowych przypadków pozwala wiarygodnie oszacować trafność predykcyjną obliczonych funkcji klasyfikacyjnych. Klasyfikacja przypadków, które posłużyły nam do obliczenia funkcji dyskryminacyjnych, może być użyteczna jedynie w identyfikowaniu przypadków odstających lub obszarów, gdzie funkcja klasyfikacyjna wydaje się być mniej trafna.
- Znaczenie podstawowych ograniczeń i wymagań metody analizy dyskryminacyjnej jest podobne jak w większości testów parametrycznych; podobne są też konsekwencje ich naruszania. Szczególnym zagrożeniem w tej metodzie są obserwacje odstające, warunkujące silne skorelowanie średnich z wariancjami, oraz przypadki niskiej tolerancji zmiennych, przyczyniające się do złego uwarunkowania macierzy danych.

Analiza log-liniowa tabel liczebności

Szczególnym przypadkiem zastosowania tablic wielopolowych, gdy chcemy testować istotność statystyczną wpływu różnych czynników, a także interakcji tych czynników, jest analiza log-liniowa (*log-linear analysis*). Zalicza się ją do metod wielowymiarowych właśnie z tego powodu, że badamy udział kilku lub kilkunastu zmiennych o rozkładach dyskretnych jednocześnie. Nazwa tej metody – analiza log-liniowa – nawiązuje do transformacji logarytmicznej danych dyskretnych, dzięki której można analizować wielopolowe tabele liczebności w kategoriach podobnych do analizy wariancji, z wyszczególnieniem efektów głównych oraz efektów interakcyjnych, które sumują się w sposób liniowy. Podobnie jak w regresji wielokrotnej, w analizie log-liniowej sprawdza się, w jaki sposób interakcje wielu zmiennych niezależnych wpływają na wartości zmiennej zależnej.

Chociaż także i w tym rodzaju analizy możemy rozróżniać zmienne zależne i niezależne – tak jak się to czyni zwyczajowo w metodzie regresji wielokrotnej lub analizie wariancji – nie jest jednak konieczne zadeklarowanie, która zmienna ma jaki charakter. O tym czy zmienna będzie zmienną objaśnianą (zależną) czy zmienną objaśniającą (niezależną), decydują wyłącznie względy racjonalne: zmienne objaśniane to te, które ulegają zmianie w reakcji na zmienne objaśniające. Na przykład, badając zależność ryzyka wystąpienia zawału od wieku pacjenta, to zdrowy rozsądek podpowiada nam, aby badać wpływ wieku (zmienna niezależna, objaśniająca) na ryzyko zawału (zmienna zależna, objaśniana), a nie odwrotnie.

Ogólna zasada tej złożonej techniki nie odbiega zasadniczo od tej, którą kierujemy się przy stosowaniu tabeli liczebności dla par zmiennych: sumaryczne liczebności brzegowe dla analizowanych czynników są odzwierciedleniem liczebności komórek, których należałoby oczekiwać, gdyby nie było zależności między tymi (dwoma lub kilkoma) czynnikami (zmiennymi). Jakiekolwiek istotne odchylenia liczebności obserwowanych od liczebności oczekiwanych wskazują na istnienie zależności między badanymi zmiennymi. W przypadku, gdy mamy do czynienia z wieloma zmiennymi, nasze poszukiwania zależności między nimi polegają na skonstruowaniu modelu, który w zadowalający sposób tłumaczyłby interakcje między różnymi czynnikami (zmiennymi). Dopasowanie takiego modelu dla zmiennych, które nie są związane, jest pod względem rachunkowym równoznaczne z obliczeniem liczebności komórek tabeli na podstawie odpowiednich liczebności brzegowych (sumarycznych). Istotne odchylenia tabeli liczebności obserwowanych od liczebności dopasowanych odzwierciedlają brak dopasowania modelu zakładającego niezależność między zmiennymi. W takiej sytuacji odrzucilibyśmy ten model i przyjęlibyśmy model, który dopuszcza zależność lub związek między badanymi zmiennymi.

Strategie doboru i dopasowania modeli w analizie log-liniowej

W metodzie analizy log-liniowej możemy badać dopasowanie różnych modeli, które odzwierciedlają różne hipotezy na temat zależności między danymi. Pierwszym naszym krokiem jest zazwyczaj zastosowanie modelu (lub modeli), który zakłada niewystępowanie jakiejkolwiek zależności między wszystkimi zmiennymi (czynnikami). Zgodnie ze stałą zasadą dla tego typu analiz, oczekiwane liczebności powinny być proporcjonalne do odpowiednich liczebności brzegowych, a wystąpienie istotnych odchyleń skłania nas do odrzucenia danego modelu. To co zapewnia nam analiza log-liniowa – a czego nie dawały prostsze metody tabel liczebności – to możliwość badania interakcji między różnymi zmiennymi w modelu. Pod tym względem metoda ta nawiązuje do metod analizy wariancji, a stosowane pojęcie interakcji jest analogiczne do tego, które stosujemy w analizie wariancji. Stwierdzenie występowania interakcji jest dla nas wskazaniem, że wpływ kilku zmiennych objaśniających wpływających na jedną zmienną objaśnianą jest wzajemnie powiązany, np. dany polimorfizm glikoprotein płytek krwi może wpływać na wystąpienie zawału, ale jedynie w określonej grupie wiekowej, albo u przedstawicieli jednej płci. Zdecydujemy wówczas, że istnieje interakcja między wpływem polimorfizmu a wpływem płci na wystąpienie epizodu zawałowego. Wraz z pojawianiem się interakcji wyższego rzędu mogą się pojawiać, tzw. modele hierarchiczne, dla których poprawne dobieranie możliwych występujących interakcji staje się coraz złożone i kłopotliwe. Ponieważ w obliczeniach metody analizy log-liniowej korzystamy rutynowo z pakietów statystycznych, w przypadku modeli hierarchicznych korzystamy najczęściej z opcji automatycznego do-

bierania możliwych interakcji oraz automatycznego dopasowywania modelu, które ułatwiają poszukiwanie modelu pasującego do danych. Ogólna logika tego algorytmu polega na dążeniu do takiego modelu końcowego, który obejmuje najmniejszą liczbę interakcji koniecznych do dopasowania naszego modelu do tabeli liczebności obserwowanych.

Podsumowanie

- Metoda analizy log-liniowej jest bardziej ogólną formą tablic wielopolowych.
- W metodzie tej badamy wpływ oraz interakcje wielu zmiennych o rozkładach dyskretnych.
- Logika obliczeń w tej metodzie nawiązuje do tej, jaką spotykamy w regresji wielokrotnej: w analizie log-liniowej sprawdza się, w jaki sposób interakcje wielu zmiennych niezależnych wpływają na wartości zmiennej zależnej.
- Z uwagi na mnogość i różnorodność dyskretnych zmiennych niezależnych, tworzymy modele obejmujące niektóre i wykluczające inne analizowane zmienne niezależne. Testując te modele orzekamy, które ze zmiennych niezależnych w modelu najlepiej opisują ten model.
- Testy istotności w metodzie analizy log-liniowej opierają się na porównywaniu par modeli oraz szacowaniu modelu lepiej dopasowanego na podstawie wartości odchyleń (równoważnych sumie kwadratów reszt w regresji wielokrotnej) między porównywanymi modelami.

Regresja logistyczna

Szczególnym przypadkiem regresji wielokrotnej są modele regresji logit i probit. Ogólnie, regresję logistyczną (*logistic regression*) opisuje równanie:

$$y = \frac{b_0}{1 + b_1 * e^{-b_2 x}},$$

które zmiennej zależnej przypisuje wartości z pewnego przedziału o ściśle wyznaczonej dolnej i górnej granicy zmiennej. Określenie tego przedziału zależy od badanego modelu doświadczalnego. Taki ogólny model regresji logistycznej jest rozwinięciem modelu logit lub logistycznego dla odpowiedzi binarnych. Osobliwą cechą modeli dla odpowiedzi binarnych jest to, że obserwowana zmienna zależna posiada rozkład dyskretny (binarny). Obliczane wartości zmiennej zależnej na podstawie równania regresji (która jest funkcją ciągłą) przyjmują wartości z zakresu od 0 do 1. W praktyce, zamiast przewidywania zmiennej binarnej, przewidujemy zmienną ciągłą, która zawiera się w granicach 0–1.
W modelu regresji logistycznej (logit), przewidywane wartości zmiennej zależnej nigdy nie są mniejsze (lub równe) od 0 ani większe (lub równe) od 1, bez względu na wartości zmiennych niezależnych. Taki rozkład zmiennej opisuje równanie regresji postaci:

$$y = \frac{e^{\left(b_0 + \sum_{i=1}^{n} b_i x_i\right)}}{1 + e^{\left(b_0 + \sum_{i=1}^{n} b_i x_i\right)}}$$

Wartość zmiennej zależnej opisywanej takim równaniem będzie zawsze oscylowała w zakresie od 0 do 1, niezależnie od współczynników regresji oraz wartości zmiennej niezależnej x. Ta użyteczna właściwość modelu logistycznego sprawia, że model ten znakomicie opisuje prawdopodobieństwo zdarzeń o rozkładzie binarnym (wystąpienie lub nie wystąpienie stanu klinicznego, wyleczenie-niewyleczenie, przeżycie-zgon, itp.). Możemy dzięki niemu oceniać, od czego zależy wielkość tego prawdopodobieństwa, to znaczy, jakie zmienne niezależne wpływają w najistotniejszy sposób na wynik zdarzenia. Metoda regresji logistycznej, służy do analizowania proporcji w sposób analogiczny jak modelu regresji wielokrotnej dla zmiennych o rozkładach ciągłych. Regresja logistyczna jest jakby pośrednią metodą w stosunku do regresji wielokrotnej i w stosunku do analizy log-liniowej, ponieważ uwzględnia zarówno zmienne ciągłe, jak i dyskretne, a także występowanie interakcji między zmiennymi (tak jak w metodzie ANOVA).

Charakterystyczną cechą funkcji logistycznej jest także jej sigmoidalny kształt, który obrazuje, że zmiany zmiennej zależnej są bardzo niewielkie zanim nie osiągną pewnej granicznej wartości progowej. Począwszy od tej wartości progowej funkcja logistyczna zaczyna gwałtownie wzrastać aż do wartości submaksymalnej (bliskiej maksymalnemu prawdopodobieństwu zdarzenia). Model logistyczny opisuje bardzo dobrze zależność prawdopodobieństwa od zmiennych objaśniających (niezależnych) dla dużych liczebności próby, tzn. 10-20-krotnie większej od liczby włączonych zmiennych niezależnych (k) (przynajmniej gdy $n > 10\ (k+1)$).

Ponieważ o rzeczywistej binarnej zmiennej zależnej y myślimy w kategoriach ukrytego ciągłego prawdopodobieństwa P z zakresu od 0 do 1, model logistyczny możemy łatwo zlinearyzować (stąd nazwa modelu – „logit" od transformacji logarytmicznej) do postaci:

$$\text{logit}\, P = \ P' = \ln \frac{P}{1-P}$$

$$\text{czyli logit} \ = \ln\left[\frac{proporcja}{1-proporcja}\right] = \frac{szansa(A)}{szansa(nieA)}$$

Po takim przekształceniu, P' może przybierać wartości z zakresu od plus do minus nieskończoności, toteż moglibyśmy zastosować wartość P' w równaniu regresji liniowej, opisanej jako:

$$\text{logit}\, P = \ P' = b_0 + \sum_{i=1}^{n} b_i x_i$$

Pod względem rachunkowym do estymacji wartości zmiennej zależnej (prawdopodobieństwa zdarzenia) wykorzystujemy technikę szacowania największej wiarygodności, w sposób analogiczny, jak stosujemy metodę najmniejszych kwadratów przy dopasowywaniu funkcji regresji liniowej. Ponieważ w metodzie regresji logistycznej operujemy prawdopodobieństwami, oznaczającymi szanse pojawienia się określonej wartości zmiennej zależnej, funkcja wiarygodności (prawdopodobieństwa) jest iloczynem prawdopodobieństw wystąpienia poszczególnych obserwacji określonych przez wiele zmiennych niezależnych w modelu. Dopasowanie modelu regresji logistycznej polega na liczeniu dla kolejnych iteracji (przybliżeń) wartości funkcji maksymalnego prawdopodobieństwa, przy założeniu, że

zmienność proporcji charakteryzuje się rozkładem dwumianowym. Rozwiązaniem modelu są wartości parametrów, dla których wiarygodność jest największa, to znaczy, przy których szansa realizacji zdarzenia jest maksymalna. Im większa jest taka wiarygodność, tym lepiej dopasowany jest nasz model regresji liniowej.

Jeżeli porównujemy prawdopodobieństwo (szansę) wystąpienia zdarzenia w różnych grupach, to wielkość wskazującą ile razy szansa zdarzenia jest większa w jednej grupie (np. grupie A) niż w drugiej (np. grupie B) nazywamy ilorazem szans (OR, *odds ratio*):

$$OR_{A\times B} = \frac{szansa(A)}{szansa(nieA)} : \frac{szansa(B)}{szansa(nieB)} = \frac{p(A)}{1-p(A)} : \frac{p(B)}{1-p(B)}$$

Iloraz szans jest wielkością bardzo często stosowaną w badaniach o charakterze epidemiologicznym. Możemy przy jego użyciu wyrażać na przykład, ile razy ryzyko wystąpienia jakiegoś czynnika ryzyka (niepomyślnego zdarzenia) jest większe w jednej grupie niż w drugiej. Wartość ilorazu szans równa w przybliżeniu 1 oznacza równoważność ryzyka w porównywanych grupach, z kolei $OR_{AxB}>1$ oznacza, że szansa niepomyślnego zdarzenia jest większa w grupie A niż w grupie B.

Pamiętamy, że wartość naszej zmiennej zależnej (dichotomicznej, 0–1) (*dichotomous variable*) zależy od wielu zmiennych (dichotomicznych, dyskretnych lub ciągłych):

$$y = P(x_1, x_2, x_3 .... x_i) = \frac{e^{\left(b_0 + \sum_{i=1}^{n} b_i x_i\right)}}{1 + e^{\left(b_0 + \sum_{i=1}^{n} b_i x_i\right)}}$$

Skoro wiemy, że

$$OR_{A\times B} = \frac{p(A)}{1-p(A)} : \frac{p(B)}{1-p(B)}$$

możemy zapisać, że

$$OR_{A\times B} = \left(\frac{e^{\left(b_0 + \sum_{i=1}^{n} b_i x_i\right)}}{1 + e^{\left(b_0 + \sum_{i=1}^{n} b_i x_i\right)}}\right)_A : \left(\frac{e^{\left(b_0 + \sum_{i=1}^{n} b_i x_i\right)}}{1 + e^{\left(b_0 + \sum_{i=1}^{n} b_i x_i\right)}}\right)_B$$

Jeżeli przyjmiemy, że wartość naszej zmiennej zależnej (np. zawału mięśnia sercowego) zależy istotnie od jednej zmiennej niezależnej (np. palenia tytoniu), to moglibyśmy zapisać:

$$OR_{A\times B} = \left(\frac{e^{(b_0 + b_i * palenie)}}{1 + e^{(b_0 + b_i * palenie)}}\right)_A : \left(\frac{e^{(b_0 + b_i * palenie)}}{1 + e^{(b_0 + b_i * palenie)}}\right)_B = \left(\frac{1}{1 + e^{-(b_0 + b_i * palenie)}}\right)_A : \left(\frac{1}{1 + e^{-(b_0 + b_i * palenie)}}\right)_B$$

$$\text{czyli} \quad OR_{A\times B} = \frac{\left(1 + e^{(b_0 + b_i * palenie)}\right)_A}{\left(1 + e^{(b_0 + b_i * palenie)}\right)_B}$$

Dla cech dichotomicznych, takich jak palenie [pali (1)-nie pali (0)] wyrażenie to da się dalej uprościć do postaci:

$$OR_{A\times B} = \frac{\left(1 + e^{b_0 + b_i * (1)}\right)_A}{\left(1 + e^{b_0 + b_i * (0)}\right)_B} = \frac{\left(e^{(b_0 + b_i)}\right)_A}{\left(e^{(b_0)}\right)_B} = e^{(b_1)}$$

Analogicznie moglibyśmy zapisać dla każdej zmiennej:

$$OR_{A\times B} = e^{b_1 * (A-B)},$$

gdzie A i B oznaczałyby wartości zmiennej w każdej z porównywanych grup.

Model probitowy (*probit logistic regression*). Jeżeli binarność zmiennej zależnej jest pozorna, to znaczy, kiedy rejestrujemy wynik jako odpowiedź binarną (tak-nie), a w rzeczywistości ta zmienna zależna posiada rozkład normalny (przyjmuje wartości ze skali o wielostopniowej gradacji), to mamy do czynienia z modelem probitowym. W modelu tym zmienna zależna o rozkładzie normalnym lub postrzeganym jako normalny, jest rejestrowana jako zmienna binarna, której przypisuje się wartość prawdopodobieństwa równego obszarowi pod krzywą rozkładu normalnego.

Rozdział 12

Metody nieparametryczne

Metody nieparametryczne należą do alternatywnych metod analizy danych numerycznych, a od metod parametrycznych odróżnia je przede wszystkim to, że nie wymagają one założeń dotyczących typu rozkładu danych, w szczególności zaś nie wymagają normalności rozkładów danych. Są one szczególnie przydatne w sytuacjach, gdy dysponujemy niewielkimi zbiorami danych o rozkładach wyraźnie odbiegających od normalności, oraz kiedy transformacja danych nie przynosi pożądanych wyników (np. w postaci normalizacji rozkładu lub jednorodności wariancji). Testy nieparametryczne stworzono w założeniu jako szybkie i proste metody weryfikacji istotności różnic, pod kątem ich zastosowania w przypadku nieznajomości parametrów rozkładu badanej zmiennej w populacji (dlatego nazywają się one metodami nieparametrycznymi, gdyż nie jest wymagana znajomość parametrów rozkładu zmiennej/zmiennych). Są one prawie tak efektywne w wykrywaniu rzeczywistych różnic jak metody parametryczne w przypadkach, gdy założenia wymagane przy stosowaniu tych ostatnich nie są naruszone, i o wiele bardziej efektywne w sytuacjach, kiedy testów parametrycznych nie możemy stosować. Ponieważ opracowano je do analizy zbiorów o małych liczebnościach, metody te są bardzo proste w rutynowym stosowaniu, gdy liczebność badanych grup nie przekracza 50 przypadków. W przypadku dużych zbiorów danych (n >100) stosowanie statystyk nieparametrycznych nie ma uzasadnienia, ponieważ gdy liczebność próby bardzo wzrasta, wówczas średnie prób podlegają rozkładowi normalnemu nawet w sytuacji, gdy odpowiednia zmienna w populacji nie posiada rozkładu normalnego. Jeżeli nadal chcemy się posługiwać w takich sytuacjach metodami nieparametrycznymi, to do liczenia statystyki testów możemy korzystać z testów proporcji. Główne zastosowanie metod nieparametrycznych to badanie istotności różnic, i z tego powodu metody te mają dwa podstawowe ograniczenia (czy wady). Po pierwsze, obliczanie odpowiednich przedziałów ufności jest bardzo skomplikowaną i rachunkowo złożoną procedurą. Po drugie, metody te nie są tak „*elastyczne*" w sytuacjach, gdy zachodzi konieczność modyfikacji testu do potrzeb konkretnego modelu doświadczalnego. Są to powody, dla których wielu badaczy stosuje te metody zupełnie wyjątkowo, starając się raczej ominąć ograniczenia, jakie towarzyszą metodom parametrycznym. Dla większości metod parametrycznych istnieją ich odpowiedniki nieparametryczne. Tą korespondencję metod parametrycznych i nieparametrycznych podaje Tabela 6. Warto nadmienić, że jako metody nieparametryczne funkcjonuje także wiele metod opartych na statystyce testu chi^2, a także ogólnych metod dotyczących proporcji lub analiz tabel wielopolowych.

Pod względem rachunkowym metody oparte są na analizie rang, czyli kolejnych numerów danych uporządkowanych monotonicznie (tzn. w kolejności rosnącej lub malejącej). W przypadkach, gdy dane mają identyczną wartość, przyporządkowujemy im tzw. rangi wiązane (*tied ranks*), liczone jako uśrednione numery kolejnych rang nadawanych kolejnym powtórzeniom tej samej wartości. Na przykład, identyczne wartości zajmujące kolejno pozycje 9, 10, 11 i 12 będą nosiły jednakową rangę (9+10+11+12)/4 = 10.5.

Tab. 6. Metody nieparametryczne.

metoda nieparametryczna	**zastosowanie**	**odpowiadająca metoda parametryczna**
test znaków	uproszczona wersja testu kolejności par Wilcoxona	
test kolejności par Wilcoxona	testowanie różnic między danymi sparowanymi	sparowany test *t*
test sumy rang Wilcoxona	porównanie dwóch grup	test *t* dla dwóch prób
test *U* Manna-Whitneya	alternatywny do testu sumy rang Wilcoxona	test *t* dla dwóch prób
jednoczynnikowa analiza wariancji Kruskala-Wallisa	porównywanie więcej niż dwóch grup	jednoczynnikowa analiza wariancji (ANOVA1)
dwuczynnikowa analiza wariancji Friedmana	porównywanie więcej niż dwóch grup; dwie zmienne grupujące	dwuczynnikowa analiza wariancji
korelacja rang Spearmana	asocjacja dwóch zmiennych	korelacja Pearsona
korelacja rang Kendalla	alternatywna do korelacji rang Spearmana	korelacja Pearsona
korelacja gamma	stosowana gdy dane zawierają wiele przypadków jednakowych rang	
test zgodności chi^2	do badania zgodności rozkładu (porównanie częstości obserwowanej z teoretyczną)	
testy Kolmogorova-Smirnova		
• dla jednej próby	alternatywny do testu zgodności chi^2	
• dla dwóch prób	porównanie dwóch rozkładów częstości	

Nieparametryczne testy istotności różnic

Testy do porównania dwóch prób

Stosujemy je, gdy porównujemy dwie grupy pod względem wartości średniej określonej zmiennej. Nieparametryczną alternatywą testu *t* dla prób niezależnych są np.: test U Manna-Whitneya, test sumy rang Wicoxona, test S Kendalla oraz test dla dwóch prób Kolmogorova-Smirnova. Jeżeli porównujemy wiele grup metodą analizy wariancji, to

stosujemy jej nieparametryczny odpowiednik – test rang Kruskala-Wallisa oraz test mediany.

Test Manna-Whitneya

Test ten zakłada, że porównywane zmienne są przynajmniej na skali porządkowej (rangowej). Różnica w stosunku do testu t dla niezależnych prób polega na tym, że obliczenia w teście U są wykonywane w oparciu o sumę rang, a nie o średnie. Wartość statystyki testu Manna-Whitneya oblicza się następująco:

$$U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - \sum R_{1j}$$

$$U' = n_2 n_1 + \frac{n_2(n_2+1)}{2} - \sum R_{2j}$$

$$U = n_1 n_2 - U'$$

Test U Manna-Whitneya jest najmocniejszym testem nieparametrycznym, w niektórych przypadkach może on nawet wykazywać większą moc przy odrzucaniu hipotezy zerowej niż parametryczny test t. Dla prób, których liczebność przekracza 20, rozkład statystyki U aproksymuje do rozkładu normalnego. Wyrazem tej dużej analogii obu testów jest podawanie obok wartości statystyki U także wartości z (wartość zmiennej o rozkładzie normalnym):

$$z = \frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 [n_1 + (n_2+1)]}{12}}}$$

lub z poprawką na rangi wiązane:

$$z = \frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 [n_1 + (n_2+1)]}{12} - \sum_{i=1}^{n} \frac{\left(t_i^3 - t_i\right)}{12}}}.$$

Hipotezę zerową odrzucamy, gdy $|z| > z_{\alpha/2}$.

Test sumy rang Wilcoxona (*Wilcoxon rank sum test*)

Analogicznie jak test Manna-Whitneya, jest nieparametrycznym odpowiednikiem testu t Studenta. Wymaga, aby porównywane zmienne były przynajmniej na skali porządkowej (rangowej). Z uwagi na duże podobieństwo do testu Manna-Whitneya (także pod względem rachunkowym) w niektórych opracowaniach opisuje się kombinowany test Manna-Whitneya-Wilcoxona:

$$U = R_2 - \frac{n_2(n_2+1)}{2}$$

$$\text{oraz} \quad U' = R_1 - \frac{n_1(n_1+1)}{2}.$$

Test mediany dla dwóch grup

Test o niewielkiej mocy (w sytuacji, gdy możemy zastosować test Manna-Whitneya lub test *t* Studenta dla dwóch prób, test mediany posiada jedynie około 67% mocy pierwszego i 64% mocy drugiego), stosowany przy niedużych liczebnościach prób zawierających (liczne) odstające obserwacje.

poziom istotności	**0.01**	**0.05**	**0.1**
test mediany	odrzucić H_0	nie odrzucać H_0	nie odrzucać H_0
test Manna-Whitneya	odrzucić H_0	odrzucić H_0	nie odrzucać H_0
test *t*	odrzucić H_0	odrzucić H_0	nie odrzucać H_0

Test mediany jest bardziej konserwatywny niż test Manna-Whitneya czy test *t* Studenta; jesteśmy też bardziej narażeni na popełnienie błędu II rodzaju.

Zgodnie z hipotezą zerową (która mówi, że wszystkie próby pochodzą z populacji o identycznych medianach) oczekujemy, że około połowa wszystkich przypadków w każdej z prób wypada powyżej, a połowa poniżej wspólnej mediany.

	grupa 1	*grupa 2*
powyżej mediany	*a*	*b*
poniżej mediany	*c*	*d*
	$n = a + b + c + d$	

przypadki „poniżej mediany" obejmują wartości ≤ Me.

Prawdopodobieństwo, że wszystkie próby pochodzą z populacji o identycznych medianach, obliczamy według równania:

$$p = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{n!a!b!c!d!}$$

Hipotezę zerową odrzucamy, gdy obliczone $p < p_\alpha$ **(mniejsze!)** dla standaryzowanego rozkładu normalnego. Test ten stosowany jest szczególnie chętnie w sytuacjach, gdy liczne przypadki znajdują się na krańcach skali pomiarowej.

Często możemy także spotkać procedurę obliczania statystyki testu mediany dla dwóch prób w oparciu o tablicę czteropolową testu chi^2:

$$\chi^2 = \frac{n\left(\left|ad-bc\right| - \frac{n}{2}\right)^2}{(a+c)(b+d)(a+b)(c+d)}, \qquad df = (k-1)(2-1) = k-1$$

W tym przypadku hipotezę zerową odrzucamy, gdy $\chi^2_{dośw} > \chi^2_{\alpha,v}$.

Test Kolmogorova-Smirnova

Przy użyciu tego testu możemy weryfikować hipotezę, że dwie próby zostały pobrane z różnych populacji. W odróżnieniu na przykład od parametrycznego testu *t* dla prób niezależnych lub testu U Manna-Whitneya, dotyczących różnic średnich lub różnic rang (czyli ogólnie – różnic w położeniu dwóch prób), test Kolmogorova-Smirnova jest także wrażliwy na różnice kształtów rozkładów w dwóch próbach (np. różnice dyspersji, skośności itd.). Ogólna zasada obliczania statystyki testu Kolmogorova-Smirnova dla jednej i dwóch prób została omówiona w *„Części II – „Badania dopasowania rozkładu"*.

Testy do porównania więcej niż dwóch prób

Test Kruskala-Wallisa

Test rang Kruskala-Wallisa (*Kruskal-Wallis test, Kruskal-Wallis one-way analysis of variance*) bada hipotezę, że próby zostały pobrane z populacji o tym samym rozkładzie lub z populacji o rozkładach posiadających tą samą medianę. Test ten zakłada, że badana zmienna ma charakter ciągły oraz że została zmierzona przynajmniej na skali porządkowej (rangowej). Jest testem analogicznym do parametrycznej jednoczynnikowej ANOVA, z tym, że jest oparty na rangach, a nie na średnich czy wariancjach.

Na podstawie rang dla *k* grup (poziomów):

grupa 1	**ranga**	**grupa 2**	**ranga**	**...**	**grupa *k***	**ranga**
d_{11}	R_{11}	d_{21}	R_{21}	...	d_{k1}	R_{k1}
d_{12}	R_{12}	d_{22}	R_{22}	...	d_{k2}	R_{k2}
d_{13}	R_{13}	d_{23}	R_{23}	...	d_{k3}	R_{k3}
...	...	...	...	...	...	...
d_{1j}	R_{1j}	d_{2j}	R_{2j}	...	d_{kj}	R_{kj}
	ΣR_{1j}		ΣR_{2j}			ΣR_{kj}

obliczamy wartość statystyki H testu Kruskala-Wallisa:

$$H = \frac{12}{N(N+1)}\left[\sum \frac{\left(\sum R_{ij}\right)^2}{n_j}\right] - 3(N+1)$$

W przypadku występowania rang wiązanych stosujemy poprawkę:

$$C = 1 - \left[\frac{\Sigma\left(t^3 - t\right)}{N^3 - N}\right], \qquad H' = \frac{H}{C}$$

Hipotezę zerową odrzucamy, gdy H (lub H') > $\chi^2_{\alpha,k-1}$.

Test mediany

Prostszą wersją testu Kruskala-Wallisa jest mniej dokładny test mediany. Obliczenia do tego testu dokonywane są w oparciu o tablice wielopolowe 2 x k i statystykę testu chi^2. Zgodnie z hipotezą zerową (która mówi, że wszystkie próby pochodzą z populacji o identycznych medianach) oczekujemy, że około połowa wszystkich przypadków w każdej z prób wypada powyżej, a połowa poniżej wspólnej mediany.

	próba 1	**próba 2**	**próba 3**	**...**	**próba *k***	**suma**
powyżej mediany	f_{11}	f_{12}	f_{13}	...	f_{1k}	R_1
nie powyżej mediany	f_{21}	f_{22}	f_{23}	...	f_{2k}	R_2
	C_1	C_2	C_3	...	C_k	n

Hipotezę zerową odrzucamy, gdy obliczona statystyka $\chi^2 > \chi^2_{\alpha,k-1}$.

Testy do porównania dwóch zmiennych

Nieparametrycznymi odpowiednikami sparowanego testu t dla grup zależnych są test znaków oraz test kolejności par Wilcoxona. Do analizy zmiennych dichotomicznych (dyskretnych) stosujemy test McNemara. Odpowiednikiem dwukierunkowej ANOVA jest dwukierunkowa analiza wariancji Friedmana, do porównywania więcej niż dwóch zmiennych z tej samej próby stosujemy test Q Cochrana (dla zmiennych dyskretnych).

Test znaków (*sign test*)

Wymaga on jedynie, aby rozkład analizowanej zmiennej był ciągły, natomiast nie jest istotne spełnienie założeń dotyczących charakterystyki rozkładu. Hipoteza zerowa (mówiąca o braku różnic między zmiennymi) zakłada, że brak różnic wystąpi w około 50%. Jeżeli para zmiennych nie różni między sobą (tak jak to zakłada hipoteza zerowa), oznacza to, że liczba rachunkowych różnic oznaczonych jako (–) jest równa liczbie różnic ze znakiem (+) (stosunek znaków wynosi 1:1 lub 50%:50%). Im bardziej zmienia się ta proporcja, tym bardziej jest istotna różnica między zmiennymi. Dla liczności <10 do obliczania statystyki testu możemy korzystać z aparatu matematycznego stosowanego do opisu rozkładu dwumianowego:

$$p(x) = \binom{n}{x} p^x q^{n-x}$$

Dla liczności prób ≥ 10, możemy zastosować test proporcji z poprawką na ciągłość Yatesa:

$$z = \frac{|p - P_0| - \frac{1}{n}}{\sqrt{\frac{(P_0)(1-P_0)}{n}}}$$

gdzie p oznacza frakcję różnic dodatnich:

$$p = \frac{(+)}{(+)+(-)}$$

Test kolejności par Wilcoxona (*Wilcoxon signed rank test*)

Test ten jest alternatywą sparowanego testu *t* Studenta. Może być stosowany we wszystkich tych sytuacjach, w których stosuje się sparowany test *t*, jednak jeżeli rozkład różnic między sparowanymi zmiennymi jest normalny, test kolejności par Wilcoxona posiada jedynie około 95% mocy testu sparowanego *t* Studenta. Test kolejności par Wilcoxona zakłada, że istnieje możliwość nadania rang wielkościom różnic par obserwacji w jednoznaczny sposób. Osobliwą cechą tego testu jest to, że część danych (tych, które dają różnicę w parach równą 0) może być usunięta.

Dla niedużych liczebności test ten można przeprowadzić w prosty sposób w czterech etapach:

- wyrzucamy różnice wynoszące zero, zaś pozostałe porządkujemy w kolejności rosnącej, zaniedbując znak różnicy; w przypadku równych wartości różnic liczymy rangi wiązane;
- sumujemy rangi dla różnic dodatnich (T_+) i te dla różnic ujemnych (T_-);
- gdyby nie było różnic między zmiennymi, to sumy T_+ oraz T_- byłyby podobne; czym większa różnica między zmiennymi, tym większa byłaby różnica między T_+ i T_-; mniejsza z obliczonych sum stanowi statystykę testu T;
- porównujemy doświadczalną wartość T z wartością krytyczną w tablicach dla testu kolejności par Wilcoxona przy istotności α oraz liczebności *N*, oznaczającej liczbę różnic różnych od zera. Jeżeli wartość doświadczalna jest **mniejsza** od krytycznej tablicowej, to odrzucamy hipotezę zerową.

Pamiętając, że w miarę wzrostu liczebności zmiennych, statystyki testów nieparametrycznych aproksymują do rozkładu normalnego, możemy także obliczyć wartość statystyki testu kolejności par Wilcoxona w alternatywny sposób. Hipoteza zerowa w tym teście zakłada, że nie ma różnic między zmiennymi, oczekiwana suma rang powinna być równo rozdzielona między dwie grupy: różnic ze znakiem (–) oraz różnic ze znakiem (+):

$$T_{oczekiwane} = \frac{n(n+1)}{2} * \frac{1}{2} = \frac{n(n+1)}{4}$$

Wartość statystyki testu wynosi:

$$z = \frac{T_{dośw} - T_{oczekiwane}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}$$

Aby odrzucić hipotezę zerową, obliczone *z* powinno być większe od tablicowego $z_{\alpha/2}$. Jeżeli wartości rang obserwowane (doświadczalne) i oczekiwane są identyczne, licznik wynosi 0 i wartość statystyki *z* również wynosi 0 – nie możemy wtedy odrzucić hipotezy zerowej i wykazać istotnej różnicy między zmiennymi.

Test McNemara

Stosujemy ten test do badania zależności lub niezależności między sparowanymi zmiennymi dyskretnymi (dichotomicznymi). Sparowane pomiary mogą dotyczyć tych samych obiektów – wtedy, gdy porównujemy zmienną w jakimś odstępie czasu:

	pomiar pierwszy	
pomiar drugi	*wynik 1*	*wynik 2*
wynik 1	*a*	*b*
wynik 2	*c*	*d*

lub też możemy porównywać dwie charakterystyki jakiejś zmiennej, np. przewidywaną i obserwowaną:

	charakterystyka 1	
charakterystyka 2	*wynik 1*	*wynik 2*
wynik 1	*a*	*b*
wynik 2	*c*	*d*

Statystykę testu liczymy w oparciu o czteropolowe tabele liczebności w następujący sposób:

$$\chi^2_{McNemara} = \frac{(b-c)^2}{b+c}$$

lub z uwzględnieniem poprawki na ciągłość Yatesa:

$$\chi^2_{McNemara} = \frac{(|b-c|-1)^2}{b+c}.$$

Dla dużych liczebności test ten jest równoważny testowi proporcji: $x^2 = z^2$, gdzie

$$z = \frac{a-d}{\sqrt{a+d}}.$$

Test Friedmana (*Friedman test, Friedman two-way analysis of variance*)

Aby posłużyć się tym testem, zmienne (w poszczególnych grupach) powinny być zmierzone przynajmniej na skali porządkowej (rangowej). Test ten odpowiada dwuczynnikowej ANOVA, a hipoteza zerowa zakłada, że porównywane zmienne zostały pobrane z populacji o równych medianach. Obliczanie statystyki polega pod względem rachunkowym na policzeniu sum rang dla k grup (poziomów) zmiennej niezależnej oraz dla n bloków (np. ochotników wybranych do badań). Ten schemat nawiązuje do metody zupełnej zrandomizowanej analizy blokowej w parametrycznej metodzie ANOVA. Statystykę testu liczymy jako:

$$\chi^2_r = \frac{12}{nk(k+1)}\sum (R_j)^2 - 3n(k+1)$$

Hipotezę zerową odrzucamy, gdy $\chi_r^2 > \chi_{\alpha,k-1}^2$. Dla niewielkich liczebności (< 5 bloków) korzystamy z tablic skonstruowanych specjalnie dla tego testu, jeżeli $n > 5$, możemy zastosować tablice testu chi^2 dla $k–1$ stopni swobody.

Test Cochrana

Test ten jest ogólniejszą wersją testu McNemara dla więcej niż dwóch prób zależnych reprezentujących liczebności lub proporcje. Przy jego użyciu weryfikujemy hipotezę, czy kilka liczebności lub proporcji istotnie różni się między sobą. Test Q Cochrana wymaga, aby zmienne były zmierzone na skali nominalnej lub zostały przetransformowane na zmienne dichotomiczne. Test ten jest odpowiednikiem testu sparowanego dla więcej niż dwóch grup dla zmiennych dyskretnych (dichotomicznych), a dokładniej – nieparametrycznym odpowiednikiem zupełniej zrandomizowanej analizy bloków dla danych zdichotomizowanych. Dla k grup dyskretnej zmiennej niezależnej:

	poziomy (grupy) zmiennej niezależnej					
	C_1	**C_2**		**C_k**	**R**	**R^2**
blok b_1	x_{11}	x_{12}	...	x_{1k}	Σx_{1k}	Σx_{1k}^2
blok b_2	x_{21}	x_{22}	...	x_{2k}	Σx_{2k}	Σx_{2k}^2
blok b_3	x_{31}	x_{32}	...	x_{3k}	Σx_{3k}	Σx_{3k}^2
....		...	...	...	...	
blok b_j	x_{j1}	x_{j2}	...	x_{jk}	Σx_{jk}	Σx_{jk}^2
C	Σx_{j1}	Σx_{j2}	...	Σx_{jk}		
C^2	Σx_{j1}^2	Σx_{j2}^2	...	Σx_{jk}^2		
				$\Sigma R =$	$\Sigma S x_k$	
				$\Sigma R^2 =$		$S x_k^2$

Wartość statystyki testu Q Cochrana wynosi:

$$Q \frac{(k-1)\left[\left(k\sum C^2\right)-\left(\sum R\right)^2\right]}{K\left(\sum R\right)-\sum R^2}, \qquad d.f. = k-1.$$

Badanie współzależności pomiędzy zmiennymi – metody korelacji nieparametrycznej (non-parametric rank correlation) oraz testy zależności zmiennych

Nieparametrycznymi odpowiednikami metody korelacji Pearsona są: korelacja R Spearmana, korelacja tau Kendalla oraz korelacja gamma. Jeżeli analizowane zmienne są zmiennymi skategoryzowanymi (dyskretnymi) (*categorized variables*), to do oceny współzależności pomiędzy zmiennymi mogą służyć na przykład test chi^2, współczynnik ϕ Cramera oraz dokładny test Fishera. W sytuacjach kiedy chcemy badać jednocześnie współzależności pomiędzy wieloma przypadkami możemy zastosować tzw. współczynnik zgodności

W Kendalla. Jako miary zależności statystycznej dla zmiennych nominalnych najczęściej wykorzystuje się takie wskaźniki jak współczynnik ϕ Cramera, współczynnik Yule'a czy współczynnik Ivesa-Gibbonsa.

Korelacja Spearmana

Współczynnik korelacji Spearmana (*Spearman's rank correlation*) posiada sens statystyczny zwyczajnego współczynnika korelacji Pearsona z tą różnicą, że oblicza się go na podstawie rang, a nie konkretnych mierzonych wartości. Dla obliczenia korelacji Spearmana zakłada się, że zmienne są mierzone co najmniej w skali porządkowej, czyli że poszczególne przypadki mogą zostać uszeregowane w dwa uporządkowane ciągi.

Współczynnik korelacji rang Spearmana (r_s, R_s) obliczamy według równania:

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)},$$

gdzie n oznacza liczbę par rang, zaś d różnicę między rangami dla każdej pary.

Istotność korelacji oceniamy porównując obliczoną wartość z wartościami krytycznymi w tablicach dla danej liczby par; jeżeli liczba par jest większa od 10, możemy także policzyć istotność w sposób analogiczny jak dla współczynnika korelacji Pearsona, tzn.:

$$t = r_s \sqrt{\left[\frac{n-2}{1-r_s^2}\right]}, \quad \text{dla } d.f. = n-2$$

Korelacja tau Kendalla

Korelacja tau Kendalla (*Kendall's rank correlation, Kendall's tau coefficient*) ma podobne założenia co test R Spearmana, a współczynniki tau i R są porównywalne co do siły wnioskowania statystycznego. Obie te miary posiadają jednak odmienną interpretację statystyczną: współczynnik R Spearmana ma sens statystyczny zwykłego współczynnika korelacji (tyle że obliczonego na podstawie rang), natomiast tau Kendalla określa prawdopodobieństwo tego, że zmierzone dane są tak samo lub różnie (odwrotnie) uszeregowane dla dwóch zmiennych:

$$\tau = \frac{2S}{N(N-1)},$$

gdzie całkowita liczba porównywanych par zmiennych wynosi

$$T = \frac{N(N-1)}{2},$$

zaś S oznacza sumę różnic punktacji zgodnych i niezgodnych kolejności dla porządku rosnącego rang (zobacz *„Część II – Uzupełnienia, przykłady i zadania"*).

Korelacja gamma

Współczynnik gamma stosujemy wtedy, gdy dane pomiarowe zawierają wiele przypadków jednakowych rang. Pod względem interpretacji i procedury obliczeniowej gamma jest bardziej podobne do tau Kendalla niż do *R* Spearmana, i reprezentuje różnicę prawdopodobieństw tego, że rangi dwóch zmiennych są ze sobą zgodne i tego, że rangi te są niezgodne, pomniejszoną o prawdopodobieństwo jednakowych rang.

Współczynniki zgodności dla zmiennych nominalnych

Współczynniki zgodności zostały opracowane z myślą o szacowaniu zależności statystycznej dla zmiennych typu nominalnego. Charakterystyczną cechą wielu takich wskaźników jest to, że wykorzystują one wartość statystyki testu chi^2. Do popularnych należą miary takie, jak na przykład:

$$C_P = \sqrt{\frac{\chi^2}{\chi^2 + n}} \qquad \text{lub} \qquad C = \sqrt{\frac{\frac{\chi^2}{n}}{1 + \frac{\chi^2}{n}}},$$

gdzie χ^2 jest wartością statystyki testu chi^2, zaś n oznacza liczbę wszystkich obserwacji (ten pierwszy wskaźnik nazywany jest nieraz także współczynnikiem Pearsona). Podstawową niewygodą ich stosowania jest to, że nie posiadają one właściwości zmian w zakresie od 0 do 1, odzwierciedlających siłę związku – czegoś, czego podświadomie oczekujemy od miary zgodności, *per analogiam* do współczynnika korelacji. Zauważmy, że każdy z tych wskaźników osiąga wartość 0 zawsze, gdy nie ma żadnej zależności między zmiennymi, to znaczy kiedy $\chi^2 = 0$. Z drugiej strony jednak, jak łatwo zauważyć – żaden ze wskaźników nigdy nie osiągnie wartości 1, nawet w przypadku, gdy istnieje całkowita zależność między zmiennymi.

Innym wskaźnikiem jest współczynnik ϕ Cramera:

$$\phi_1 = \sqrt{\frac{\chi^2}{n}},$$

który w wersji dla podwójnej dichotomii, czyli dla tablicy czteropolowej, nazywany jest także korelacją czteropunktową. Dla zmiennych mających więcej niż dwa poziomy (dwie grupy) można obliczyć ten współczynnik na podstawie wartości statystyki testu chi^2 dla tablicy wielopolowej:

$$\phi_1' = \sqrt{\frac{\chi^2}{n(k-1)}},$$

gdzie χ^2 jest wartością statystyki testu chi^2 (nie skorygowaną na ciągłość), n oznacza liczbę wszystkich obserwacji, zaś k jest liczbą rzędów lub kolumn (która jest mniejsza). W tej formie, wartość tego wskaźnika zmienia się w zakresie od 0 do 1.

Zauważmy jednak, że gdy zastosujemy wartość χ^2 obliczoną z rozkładu tablicy czteropolowej:

	jest	*brak*	
jest	*a*	*b*	***a+b***
brak	*c*	*d*	***c+d***
	a+c	***b+d***	***n***

równą:

$$\chi^2 = \frac{n[(a)(d)-(b)(c)]^2}{(a+b)(c+d)(a+c)(b+d)},$$

to współczynnik Cramera możemy także zapisać jako:

$$\phi_2 = \frac{(a)(d)-(b)(c)}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}.$$

Taki zapis ma tę zaletę, że liczony współczynnik Cramera zmienia się w zakresie od –1 do +1 i może wskazywać także kierunek zależności (wartość 0 oznacza brak zależności). Zatem jego sens statystyczny jest podobny jak zwyczajnego współczynnika korelacji liniowej. Zwróćmy uwagę, że zależność możemy tutaj łatwo interpretować w kategoriach zgodności: duża zgodność (czyli niewielka zależność między zmiennymi) wyrażająca się wyższą wartością współczynnika ϕ oznacza większe podobieństwo proporcji (liczebności w komórkach tablicy czteropolowej) dla porównywanych poziomów zmiennych dyskretnych.

Współczynnik Yule'a, równy:

$$Q = \frac{(a)(d)-(b)(c)}{(a)(d)+(b)(c)},$$

zmienia się w zakresie od –1 (wtedy gdy *a* lub *d* wynosi 0) do +1 (gdy *b* lub *c* wynosi 0), podobnie jak w przypadku współczynnika Ivesa-Gibbonsa:

$$r_n = \frac{(a+d)-(b+c)}{(a+d)+(b+c)},$$

którego sens statystyczny jest niemal identyczny jak dla współczynnika ϕ_2 Cramera.

Regresja nieparametryczna

W przypadkach, gdy nie jest spełniony warunek normalnego rozkładu zmiennej zależnej, i nie możemy posłużyć się metodą liniowej regresji w celu wyznaczenia współczynników równania regresji, możemy *„ominąć"* ten problem stosując procedury nieparametryczne. Najprostszą z nich jest tzw. *„niezupełna"* metoda Theila (*„incomplete" Theil's method*). Dla każdej pary punktów, przy spełnieniu warunku $x_j > x_i$, można wyznaczyć współczynnik kierunkowy prostej łączącej dwa wybrane punkty, b_{ij}. Tak policzone współ-

czynniki *b* są następnie uporządkowane w kolejności rosnącej w celu wyznaczenia mediany współczynnika *b*. Dla każdej pary punktów danych cząstkowych wyznacza się następnie współczynnik *a* (rzędną zerową). „*Całościowy*" współczynnik *a* (dla całego zbioru par punktów) jest równy medianie oszacowanych cząstkowych *a*.

Podstawowe zalety metody Theila to: (1) metoda nie wymaga normalności rozkładów zmiennych *x* i *y*, (2) metoda nie zakłada nierównocenności błędów *x* i *y*, (3) nie jest wrażliwa na obserwacje odstające, ponieważ nie jest oparta na rachunku sumy kwadratów różnic.

Podsumowanie

- Wiele metod nieparametrycznych to szybkie i proste metody weryfikacji istotności różnic, pod kątem ich zastosowania w przypadku nieznajomości parametrów rozkładu w populacji badanej zmiennej.
- Metody nieparametryczne nie wymagają spełniania założeń (dotyczących typu rozkładu danych, w szczególności nie wymagają normalności rozkładów danych) i nie podlegają ograniczeniom testów parametrycznych.
- Są one szczególnie przydatne w sytuacjach, gdy dysponujemy niewielkimi zbiorami danych o rozkładach wyraźnie odbiegających od normalności, kiedy transformacja danych nie przynosi pożądanych wyników (np. w postaci normalizacji rozkładu lub jednorodności wariancji).
- Metody nieparametryczne są prawie tak efektywne (mają porównywalną moc statystyczną) w wykrywaniu rzeczywistych różnic jak metody parametryczne w przypadkach, gdy założenia wymagane przy stosowaniu tych ostatnich nie są naruszone, i o wiele bardziej efektywne w sytuacjach, kiedy testów parametrycznych nie możemy stosować.
- Metody nieparametryczne służą głównie do analizy zbiorów o małych liczebnościach.
- Dla dużych liczebności nieparametryczne testy istotności różnic aproksymują do testu normalnego lub testu *t* Studenta (dla danych ciągłych) lub testów proporcji (dla danych dyskretnych).
- Korzystając z metod nieparametrycznych należy starannie dobierać te, które charakteryzują się jak największą mocą statystyczną.
- Moc testu silnie zależy od liczebności próby (lub liczebności w poszczególnych kategoriach).
- Testy nieparametryczne górują mocą nad metodami parametrycznymi przy małych liczebnościach próby.

Rozdział 13

Metody wykorzystywane w badaniach populacyjnych i diagnostycznych

Miary zapadalności i umieralności

W najprostszym przypadku miary zapadalności (*measures of morbidity*) lub umieralności (*measures of mortality*) szacowane są jako stosunek liczebności przypadków zachorowań lub śmierci do całkowitej liczebności populacji poddanej obserwacji. Sytuacja komplikuje się kiedy mamy do czynienia na przykład ze stratyfikacją wiekową lub w przypadkach standaryzacji struktury wieku, płci, itp. badanych grup.

Wskaźniki urodzeń i umieralności

Wyraża się je na rok na 1000 osobników. Najczęściej stosowane miary to:

1. wskaźnik urodzeń (*birth rate*) =

$$= \frac{liczba\ urodzeń\ na\ rok}{liczebność\ \ populacji} \times 1000$$

2. wskaźnik płodności (*fertility rate*) =

$$= \frac{liczba\ \ urodzeń\ \dot{z}ywych\ noworodków\ na\ \ rok}{liczebność\ \ \ populacji\ \ kobiet\ \ w\ \ wieku\ \ 15-49\ \ lat} \times 1000$$

3. współczynnik umieralności ogólnej (*crude mortality rate*) =

$$= \frac{liczba\ zgonów\ ogółem\ w\ danym\ czasie}{liczba\ ludności\ narażona\ na\ ryzyko\ zgonu\ w\ danym\ czasie} \times\ 1000$$

4. cząstkowe współczynniki umieralności (*partial mortality rates*), np.:
współczynnik umieralności w grupie wiekowej (ze stratyfikacją) (*age-specific mortality rate/ stratified mortality rate*) =

$$= \frac{\textit{liczba zgonów w grupie wiekowej}}{\textit{liczba ludności w grupie wiekowej narażona na ryzyko zgonu}} \times 1000$$

współczynnik umieralności w grupie mężczyzn =

$$= \frac{\textit{liczba zgonów w grupie mężczyzn}}{\textit{liczba mężczyzn narażona na ryzyko zgonu}} \times 1000$$

5. wskaźnik umieralności proporcjonalnej (*proportional mortality rate*) jest proporcją zgonów z powodu określonej przyczyny w stosunku do ogólnej liczby zgonów, np. z powodu za krzepicy żył głębokich wskaźnik taki wynosi:

$$= \frac{\textit{liczba zgonów z powodu zakrzepicy żył głębokich}}{\textit{liczba zgonów ogółem}} \times 100$$

Od współczynników umieralności należy odróżnić wskaźnik śmiertelności, który wyraża, jaka jest proporcja zgonów na określoną chorobę w stosunku do ogólnej liczby chorych:

6. śmiertelność ogólna =

$$= \frac{\textit{liczba zgonów na daną chorobę}}{\textit{liczba chorych ogółem}} \times 100$$

7. śmiertelność chorobowa (*case fatality rate*) =

$$= \frac{\textit{liczba zgonów na rok z powodu określonej choroby}}{\textit{liczba przypadków choroby ogółem}} \times 100$$

Zwróćmy uwagę, że wyrażenie mianownika w powyższych równaniach obejmuje wszystkie przypadki osób chorych, także te, które występowały w badanej populacji zanim przystąpiliśmy do badania.

Wskaźniki urodzeń, umieralności ogólnej i cząstkowej uważa się za prawdziwe miary częstości, które wyrażają jak szybko populacja zmienia się w czasie (np. jak szybko jej liczebność wzrasta lub maleje). W praktyce za liczebność populacji (wyrażenie mianownika) przyjmuje się wartość rejestrowaną w środkowym przedziale obserwacji (czyli jeżeli np. przedział czasu prowadzenia obserwacji obejmuje rok, to wartość taką rejestrowalibyśmy w 183 dniu obserwacji).

Obok tego mamy do dyspozycji miary ryzyka, określające jakie jest prawdopodobieństwo wystąpienia określonego zdarzenia, np. ryzyko śmierci w przypadku wystąpienia określonej choroby. Do takich miar ryzyka zaliczymy na przykład współczynniki umieralności niemowląt. W stosowanej tutaj terminologii należy przyjąć, że:

- zgon płodu to taki zgon, który nastąpił przed jego wydaleniem lub usunięciem z ciała matki;
- noworodkiem martwo urodzonym będzie taki noworodek, którego zgon nastąpił przed jego wydaleniem lub usunięciem z ciała matki i którego masa w chwili urodzenia wynosiła co najmniej 1001 g.

Do najważniejszych współczynników w ocenie umieralności niemowląt należą:

8. ogólny współczynnik umieralności niemowląt (*infant mortality rate*) =

$$= \frac{\textit{liczba zgonów niemowląt 0-11 miesięcy w danym roku}}{\textit{liczba urodzeń żywych w danym roku}} \times 1000$$

9. współczynnik wczesnej umieralności niemowląt (*neonatal mortality rate*) =

$$= \frac{\textit{liczba zgonów niemowląt w wieku 0–27 dni w danym roku}}{\textit{liczba urodzeń żywych w danym roku}} \times 1000$$

10. współczynnik późnej umieralności niemowląt =

$$= \frac{\textit{liczba zgonów niemowląt w wieku 28 dni–11 miesięcy w danym roku}}{\textit{liczba urodzeń żywych w danym roku}} \times 1000$$

11. współczynnik umieralności okołoporodowej niemowląt (*perinatal mortality rate*) =

$$= \frac{\textit{liczba urodzeń martwych+liczba zgonów niem. w wieku <7 dni w danym roku}}{\textit{liczba urodzeń (martwych + żywych) w danym roku}} \times 1000$$

Chociaż w definicji umieralności uwzględniamy liczbę przypadków śmierci rejestrowanych w okresie 1 roku, w praktyce rzadko mamy możliwość prowadzić obserwacje poszczególnych przypadków (osób, pacjentów) dokładnie w tym samym przedziale czasu w okresie jednego roku. Składa się na to wiele przyczyn, np. migracje ludności, wyłączenie z badań lub włączenie nowych ciekawych przypadków, nieprzewidywalność pojawiania się nowych przypadków choroby, itp. Toteż często zdarza się, że obserwacje wybranych losowo przypadków prowadzimy w okresie o wiele krótszym niż jeden rok. Z tych względów wygodniejszym do stosowania w praktyce wskaźnikiem jest umieralność zdefiniowana jako:

$$\text{umieralność} = \frac{\textit{liczba zgonów}}{\textit{liczba osobo–lat w okresie obserwacji}} \times 1000$$

Przy takim oszacowaniu liczba osobo-lat (*person-years*) będzie taka sama dla jednej osoby obserwowanej przez okres 1 roku jak dla 12 osób, z których każda będzie obserwowana przez 1 miesiąc. Z uwagi na mnożnik 1000 takiego wskaźnika, powinniśmy zwracać uwagę na sposób wyrażania wyników: na przykład należy rozróżniać między miarą na 1000 osób na rok a miarą na 1000 osobo-lat.

Typowym przykładem badania obserwacyjnego umieralności w jakiejś społeczności jest rejestrowanie liczebności badanej grupy dwukrotnie i obliczanie liczby przypadków śmierci. W takim badaniu osoba rejestrowana w obu badaniach przeprowadzanych w przy-

jętym odstępie czasu stanowi składową osobo-lat w całym okresie prowadzenia obserwacji w wyrażeniu mianownika. Osoba, która została wyłączona z badania (np. dlatego że migrowała lub zmarła) jest składową osobo-lat w okresie od pierwszej rejestracji do daty śmierci lub migracji, podobnie osoba, która urodziła się w czasie trwania badania jest składową osobo-lat w okresie od daty urodzin do czasu drugiej rejestracji.

Zapadalność i chorobowość

Są to dwie zasadnicze miary częstości występowania/wystąpienia i rozpowszechnienia danej choroby, którą można zaklasyfikować do jednej z czterech grup:

- chorób, które zaczęły się i zakończyły w okresie prowadzenia obserwacji,
- chorób, które zaczęły się w okresie prowadzenia obserwacji, ale trwają nadal,
- chorób, które zaczęły się dawniej, ale zakończyły się w okresie prowadzenia obserwacji,
- chorób, które zaczęły się dawniej i trwają nadal.

Zapadalność (*incidence*, zachorowalność, częstość wystąpienia choroby) oznacza liczbę nowych przypadków wystąpienia choroby w okresie prowadzenia obserwacji w stosunku do liczby osób, które potencjalnie mogłyby zachorować.

Może być zdefiniowana albo w kategoriach ryzyka wystąpienia choroby (*incidence risk*) albo częstości wystąpienia choroby (*incidence rate*). Zapadalność definiowana jako ryzyko wystąpienia choroby (*incidence risk*) to prawdopodobieństwo, że osoba pierwotnie zdrowa zachoruje w jakimś momencie prowadzenia obserwacji. Częstość wystąpienia choroby (*incidence rate*) oznacza natomiast liczbę osób, które zachorują do liczby wszystkich osób potencjalnie narażonych na zachorowanie (czyli takich którzy jeszcze ciągle są zdrowi). W takim rozumieniu, osoba, która zapada na chorobę nie jest już dłużej osobą potencjalnie narażoną na zachorowanie. Liczba nowych przypadków zachorowań odnoszona jest zatem nie do liczby osób, które były narażone na początku badania, ale do iloczynu uśrednionej liczby osób narażonych w okresie prowadzenia obserwacji i czasu trwania obserwacji (tzw. osobo-lata narażone na zachorowanie, *person-years at risk*, pyar). Miara osobo-lat narażonych na zachorowanie jest analogiczna jak w przypadku osobo-lat w okresie obserwacji, z wyjątkiem powszechnych chorób, na które można zapaść więcej niż raz w okresie obserwacji (przeziębienie, biegunka, itp.): zamiast czasu od rozpoczęcia badania do zachorowania stosuje się sumaryczny czas kiedy nie występuje choroba. W przypadku chorób powszechnie występujących, na które można zapadać często częstość wystąpienia choroby określa średnią liczbę ataków choroby na osobę (narażoną czyli nie chorującą) na rok:

$$\text{zapadalność } (\textit{ryzyko wystąpienia choroby}) = \frac{\textit{liczba nowych zachorowań w okresie obserwacji}}{\textit{liczba osób narażonych na początku obserwacji}}$$

$$\text{zapadalność } (\textit{częstość choroby}) = \frac{\textit{liczba nowych zachorowań w okresie obserwacji}}{\textit{liczba osobo–lat narażonych w okresie obserwacji}} =$$

$$= \frac{\textit{liczba nowych zachorowań w okresie obserwacji}}{(\textit{średnia liczba narażonych w okresie obserwacji}) \times (\textit{czas obserwacji})}$$

Te obie miary są niemal jednakowe w przypadku rzadkich chorób (np. niektóre nowotwory) lub chorób przewlekłych, natomiast wyraźnie różne w przypadku chorób występujących powszechnie (lub chorób o krótkim przebiegu).

Stosunek tych obu miar jest nazywany ryzykiem względnym (*zobacz niżej*).

Chorobowość (ang. *prevalence*) natomiast to całkowita liczba przypadków choroby już występujących w danej populacji (bez względu na to kiedy zachorowały). Chorobowość zależy oczywiście zarówno od zapadalności (zachorowalności), jak i od czasu trwania choroby. Przy określonej zapadalności chorobowość będzie tym większa, im choroba będzie bardziej długotrwała.

Chorobowość można rejestrować w danym konkretnym momencie (*chorobowość punktowa*), jak i w dłuższym okresie prowadzenia obserwacji (*chorobowość okresowa*) (miary te wyraża się w %):

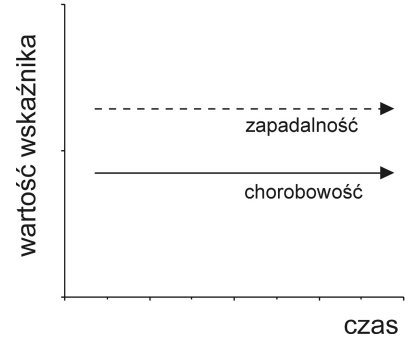
$$\text{chorobowość punktowa} = \frac{\textit{liczba chorych osób w danym momencie}}{\textit{liczebność populacji}}$$

$$\text{chorobowość okresowa} = \frac{\textit{liczba chorych osób w okresie obserwacji}}{\textit{liczebność populacji w punkcie środkowym}}$$

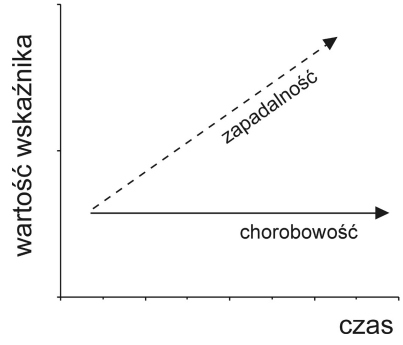
Ogólnie, zapadalność dotyczy jedynie nowych zachorowań, natomiast chorobowość zarówno nowych jak i zadawnionych. Obliczając wartość tych współczynników należy sprecyzować, czy wyrażenie w liczniku dotyczy liczby chorych osób czy liczby epizodów choroby. Zapadalność i chorobowość połączone są zależnością:

chorobowość = zapadalność x czas trwania choroby

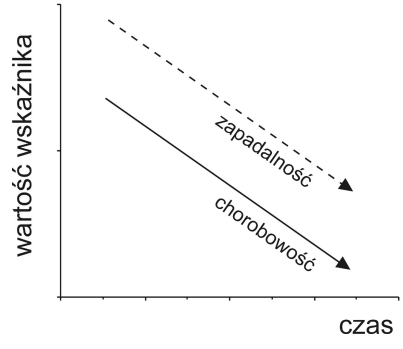
Jeżeli żaden ze współczynników nie ulega zmianom w czasie to znaczy, że śmiertelność i wyleczalność danej choroby są stałe (może tak być jeżeli nie występuje migracja osób zdrowych i osób chorych).



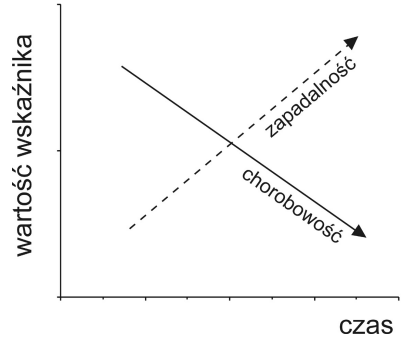
Jeżeli zapadalność wzrasta w jakimś okresie prowadzenia obserwacji, a chorobowość nie zmienia się, oznacza to, że wyleczalność lub śmiertelność rośnie (w wyniku tego czas trwania choroby skraca się). Może tak być na przykład kiedy profilaktyka nie odnosi skutku, natomiast skuteczność leczenia poprawia się.



Gdy oba współczynniki zmniejszają się w czasie proporcjonalnie do siebie, może to odzwierciedlać sytuację, kiedy zmniejszenie zapadalności, związane na przykład z obniżeniem poziomu czynników ryzyka w badanej populacji (w wyniku skutecznej profilaktyki), idzie w parze ze zmniejszeniem chorobowości (na skutek polepszenia wyleczalności).



Jeżeli zmiany obu współczynników są przeciwstawne, na przykład zapadalność rośnie a chorobowość zmniejsza się (tak jak na diagramie poniżej), obrazuje to przypadki, gdy prewencja kliniczna (skuteczność leczenia choroby) poprawia się, zaś liczba nowych przypadków zachorowań wzrasta (np. w związku ze wzrostem nasilenia czynników ryzyka: zmianą sposobu życia, stresem, zmianą odżywiania, itp.).



Analiza częstości wystąpienia choroby z wykorzystaniem rozkładu Poissona

Jak pamiętamy z Rozdziału „*Rozkład Poissona*", charakterystyka tego rozkładu może być wykorzystana do analizy częstości zdarzeń. Szczególnym przypadkiem takiej analizy jest analiza częstości wystąpienia choroby (*incidence rate*). Pamiętamy, że częstość wystąpienia choroby to liczba nowych przypadków choroby (x) zaobserwowana w przedziale czasu podzielona przez całkowitą liczbę osobo-lat narażonych na zachorowanie (*person-years at risk*, pyar). Jeżeli możemy założyć, że obserwowane zdarzenia zachorowań są niezależne od siebie i mają przypadkowy rozkład w czasie (niestety w rzeczywistości często założenie to nie jest spełnione, np. w chorobach zakaźnych), to w analizie takiej możemy wykorzystać rozkład Poissona oraz jego normalną aproksymację:

$$\lambda = \frac{x}{pyar}, \qquad SE = \frac{\sqrt{x}}{pyar} = \sqrt{\frac{\lambda}{pyar}}.$$

Korzystając z aproksymacji normalnej rozkładu Poissona, możemy porównywać dwie częstości wystąpienia choroby $\lambda_1 = x_1 / pyar_1$ oraz $\lambda_2 = x_2 / pyar_2$ w następujący sposób (równanie z poprawką na ciągłość):

$$z = \frac{\left|\lambda_1 - \lambda_2\right| - \left[1/(2\,pyar_1) + 1/(2\,pyar_2)\right]}{\sqrt{\left[\lambda(1/pyar_1 + 1/pyar_2)\right]}},$$

gdzie $\lambda = \frac{x_1 + x_2}{pyar_1 + pyar_2}$ oznacza całkowitą częstość wystąpienia choroby w obu porównywanych grupach łącznie.

Standaryzacja danych bezpośrednia i pośrednia

Ponieważ zapadalność i chorobowość są zwykle silnie zależne od struktury wiekowej, a także proporcji płci w badanych populacjach, aby w pełni wiarygodnie porównywać te miary w różnych badanych populacjach, należy najpierw dokonać standaryzacji miar w odniesieniu do płci i wieku w porównywanych grupach. Wydaje się to oczywiste, jeżeli zauważymy na przykład, że umieralność w populacji starszej będzie naturalnie większa niż w populacji młodej, co rzecz jasna nie uprawnia nas do stwierdzenia, iż dowolnie wylosowana osoba z populacji, gdzie średnia wieku jest wyższa, będzie narażona na śmierć bardziej niż dowolnie wylosowana osoba z populacji o niższym średnim wieku.

Do standaryzacji możemy stosować metody bezpośrednie lub pośrednie (*direct*/*indirect standardization*), w zależności od tego, w jaki sposób dobieramy grupę standardową. Taką grupę może stanowić wybrana przez nas jedna z porównywanych populacji, może to być populacja złożona z obu populacji porównywanych, lub populacja określonego regionu czy grupy ludzi. Nawet przy takiej dużej arbitralności doboru populacji standardowej, ryzyko niepoprawnej interpretacji wyników oraz błędnych wniosków jest niewielkie, mimo, że wartości bezwzględne miar mogą się różnić w zależności od obranej metody. Ogólnie,

metody pośrednie są stosowane częściej przy szacowaniu miar umieralności i zapadalności, natomiast bezpośrednie przy analizie chorobowości.

Metody standaryzacji możemy stosować niezależnie od tego czy miary umieralności lub zachorowalności oceniamy w kategoriach ryzyka czy częstości wystąpienia. Alternatywnie, możemy analizować te miary stosując testy proporcji, statystykę rozkładu Poissona lub modele log-liniowe, pod warunkiem, że zdarzenia (zgony, zachorowania, itp.) występują niezależnie od siebie i są równomiernie rozłożone w czasie.

Procedury standaryzacji omówione zostały dokładniej w *„Części II – Uzupełnienia, przykłady i zadania"*.

Metody oparte na statystyce testu chi^2

Według konwencjonalnego podziału na retrospektywne i prospektywne (z ciągłym monitorowaniem) badania populacyjne, możemy wyróżnić dwie strategie algorytmu badawczego. Badania retrospektywne to takie, w których różnicujemy jak określona jednostka chorobowa (która pojawiła się kiedyś w przeszłości) wpływa na częstość wystąpienia interesujących nas modyfikacji zmiennej. W badaniach tego typu spoglądamy zatem wstecz i oceniamy jak status kliniczny – który możemy traktować jako rodzaj modulatora – wpływa na wartości zmiennej (lub zmiennych). Badamy ryzyko tego, że wystąpienie tego stanu klinicznego zmieni wartości badanych zmiennych. Schemat postępowania w takich badaniach wygląda następująco:

etap 1: wybierz

		przypadki	*kontrole*
etap 2: zmierz ekspozycję w przeszłości	*ekspozycja* (+)	***a***	***b***
	ekspozycja (–)	***c***	***d***
	ogółem	***a+c***	***b+d***

etap 3: porównaj proporcje w grupach poddanych ekspozycji — ***a/(a+c)*** *względem* ***b/(b+d)***

etap 4: oblicz iloraz szans, czyli miarę wskazującą ile razy szansa ekspozycji w grupie przypadków jest różna od szansy ekspozycji w grupie kontroli — ***OR = a/c : b/d***

etap 5: oblicz proporcję przypadków przypisaną ekspozycji — ***PC = [p*(OR-1)]/[(p*(OR-1))+1]***

* *p* oznacza proporcję populacji narażoną na działanie czynnika

Naszą intencją prowadzenia takiej obserwacji jest chęć przekonania się, czy badane zmienne mogłyby dla nas być w przyszłości predyktorami wystąpienia choroby, tzn. czy sama obserwacja zmian wartości określonych parametrów może być dla nas wskazówką, że występuje określona jednostka chorobowa. Abyśmy mieli jakąś miarę odniesienia musimy prowadzić równoległe obserwacje w grupie z interesującą nas jednostką chorobową oraz w grupie referencyjnej osób bez choroby. Z tego powodu badania tego typu nazywa się badaniami kliniczno-kontrolnymi (*case-control*), ponieważ wartości zmiennej/zmiennych

u pacjentów (*cases*, kliniczne przypadki chorobowe) porównujemy z wartościami tych zmiennych w osób zdrowych (*control*). Relację, która opisuje jaki jest wpływ badanego stanu klinicznego na wartości zmiennej, wystąpienie zmian lub przebieg procesu, wyrażamy matematycznie w postaci tak zwanego ***ilorazu szans*** (*odds ratio*, OR) .

Badania prospektywne to takie badania, w których śledzimy zachowanie się lub wystąpienie zmian określonych zmiennych po (jednokrotnym lub wielokrotnym) zadziałaniu czynnika modyfikującego (modulatora). Badania te mają formę aktywnej (badacz manipuluje działaniem czynnika) i dynamicznej (monitorowanie odbywa się w sposób ciągły) obserwacji. W badaniu prospektywnym plan postępowania wygląda następująco:

etap 1: wybierz

	osoby narażone na działanie czynnika	*osoby nie narażone na działanie czynnika*
etap 2: obserwuj stan zdrowia osób w obu grupach		
nowe zachorowania	***a***	***b***
zdrowi	***c***	***d***
ogółem	***a + c***	***b + d***

etap 3: porównaj proporcje chorych w obu grupach (*określ zapadalność*)

$Z_{eksp} = a/(a+c)$ *względem* $Z_{nieeksp} = b/(b+d)$

etap 4: oblicz ryzyko względne RR = $Z_{eksp}/Z_{nieeksp}$, czyli miarę wskazującą ile razy zapadalność w grupie eksponowanych jest różna od zapadalności w grupie nieeksponowanych (kontrolnej)

$RR = Z_{eksp}/Z_{nieeksp}$

etap 5: oblicz ryzyko przypisane ekspozycji

$AR = Z_{eksp} - Z_{nieeksp}$

Taki dynamiczny wpływ badanego czynnika na interesujące nas parametry populacji wyrażamy jako współczynnik ***ryzyka względnego***.

Ryzyko względne (RR, *relative ratio*) to iloraz ryzyka wyniku dodatniego (zmiany) po zadziałaniu czynnika oraz ryzyka wystąpienia takiego wyniku bez zadziałania czynnika, czyli:

$$RR = \frac{\textit{częstość występowania zmiany po zadziałaniu czynnika}}{\textit{częstość występowania zmiany bez zadziałania czynnika}}$$

To co określamy tutaj jako częstość wystąpienia zmiany można także określić jako częstość występowania choroby lub zapadalność (*incidence*), ponieważ bardzo często przedmiotem analizy są czynniki ryzyka zwiększające prawdopodobieństwo jakiegoś procesu patologicznego, np. wpływ tytoniu na zapadalność na raka płuc, czy wpływ cholesterolu na zawał mięśnia sercowego. Zwróćmy uwagę, że terminy „ryzyko" oraz „częstość" są w tych definicjach stosowane wymiennie; nie dlatego, że są one sobie tożsame, ale ponieważ konsekwencją większego ryzyka wystąpienia choroby w populacji jest większa szansa że zachoruje więcej ludzi, a tym samym częstość choroby będzie większa. Jeżeli ryzyko (rozumiane jako częstość) wystąpienia zmian lub zmiany (np. choroby) po zadziałaniu czynnika jest takie samo w obu grupach: badanej i kontrolnej, to oczywiście RR będzie

równy jeden, co oznacza, że nie ma związku przyczynowo-skutkowego między występowaniem (działaniem) czynnika a zapadalnością. Współczynnik RR większy od 1 występuje gdy ryzyko wystąpienia zmian (zachorowania) jest większe w grupie poddanej działaniu czynnika; wartości istotnie wyższe od 1 wskazują, że czynnik może być uważany za czynnik ryzyka. Wartości RR niższe od 1 sugerują, że czynnik działa ochronnie. Czym bardziej wartość RR odbiega od 1, tym większa jest zależność badanego czynnika z występowaniem zmian (zapadalnością).

RR obliczamy w następujący sposób z liczności w komórkach tablicy czteropolowej:

		obecność czynnika		
		występuje	*nie występuje*	
wynik	*dodatni*	*a*	*b*	***a+b***
	ujemny	*c*	*d*	***c+d***
		a+c	***b+d***	***n***

Stąd mamy:

ryzyko wyniku (dodatniego) po zadziałaniu czynnika = $\frac{a}{a+c}$

ryzyko wyniku (dodatniego) bez zadziałania czynnika = $\frac{b}{b+d}$

ryzyko względne (RR) = $\frac{\frac{a}{a+c}}{\frac{b}{b+d}} = \frac{ab+ad}{ab+bc}$

Istotność ryzyka względnego jest obliczana na podstawie wartości odpowiedniej statystyki chi^2:

$$\chi^2 = \frac{n(ad-bc)^2}{(a+b)(b+d)(a+b)(c+d)}$$

Możemy także oszacować przedział ufności dla współczynnika RR:

dla 95% przedziału ufności $95\%\ CI = RR^{(1\pm 1.96/\chi)}$

lub dla jakiegokolwiek przedziału ufności $CI = RR^{(1\pm z/\chi)}$.

Aby wyrazić o ile – a nie ile razy – wzrośnie ryzyko wystąpienia zmiany (zachorowania) stosujemy parametr zwany zmianą względnego ryzyka lub ***ryzykiem przypisanym*** (*attributable risk, AR*):

AR = (częstość występowania zmiany po zadziałaniu czynnika) – (częstość występowania zmiany bez zadziałania czynnika)

Jeżeli zmiana ryzyka względnego będzie wyrażona jako proporcja w stosunku do częstości występowania zmiany po zadziałaniu czynnika, to mamy do czynienia z tzw. ***proporcjonalnym ryzykiem przypisanym*** (proporcjonalną zmianą ryzyka względnego, *proportional attributable risk, PAR*):

$$PAR = \frac{RR-1}{RR}$$

Sens statystyczny PAR jest też taki, że wskazuje on za jaką frakcję wszystkich przypadków wyniku pozytywnego odpowiada zadziałanie badanego czynnika.

Pomimo, że zmiana względnego ryzyka (AR) daje najlepsze pojęcie o wzroście ryzyka zachorowalności po wystawieniu na działanie czynnika, całościowy wpływ czynnika na populację zależy także od tego, jaka jest częstość występowania tego czynnika w otoczeniu. Rzadkie i okazjonalne ekspozycje na działanie czynnika mają niewielki całościowy efekt nawet wtedy, gdy czynnik zwiększa bardzo istotnie ryzyko zachorowalności. W takim przypadku podajemy wartość ***ryzyka przypisanego w populacji*** (*overall attributable risk, OAR*):

OAR = (całościowa częstość występowania zmiany po zadziałaniu czynnika) – (częstość występowania zmiany bez zadziałania czynnika)

Odpowiedni wskaźnik proporcjonalny (POAR, *proportional overall attributable risk*) jest ilorazem: (całościowa zapadalność – zapadalność bez zadziałania czynnika)/(całościowa zapadalność). Wyrażamy go jako:

$$\text{POAR} = \frac{chorobowość_{ekspozycja}\,(RR-1)}{1+chorobowość_{ekspozycja}\,(RR-1)}$$

Alternatywną miarą zapadalności może być stosunek zachorowań do niezachorowań (szansa zachorowania), rachunkowo równy ilorazowi liczby przypadków, które zachorowały, do liczby tych, które nie zachorowały w czasie trwania obserwacji, ale przy końcu badania istnieje nadal ryzyko, że zachorują. Proporcja szansy zachorowania wśród osób poddanych działaniu czynnika (a/b) oraz szansy zachorowania wśród osób nie poddanych działaniu czynnika (c/d), nazwana jest właśnie ***ilorazem szans*** (OR, *odds ratio*):

	choroba	*brak choroby*	
wystawienie na działanie czynnika	*a*	*b*	***e***
nie wystawienie na działanie czynnika	*c*	*d*	***f***
	g	***h***	***n***

$$OR = \frac{a/b}{c/d} = \frac{ad}{bc}$$

Równoważnie, OR można traktować jako stosunek szans a/c (narażona lub nie narażona na działanie czynnika) wśród osób chorych do analogicznej szansy u osób zdrowych, ponieważ:

$$OR = \frac{a/c}{b/d} = \frac{ad}{bc}$$

Ze względu na tą właściwość „przemienności" miara ta jest tak użyteczna w badaniach kliniczno-kontrolnych (*case-control*).

Analogicznie, jak dla współczynnika ryzyka względnego, przedział ufności OR obliczamy jako:

$$CI(\%) = OR^{(1\pm z/\chi)} \quad \text{oraz} \quad 95\%CI = OR^{(1\pm 1.96/\chi)} \quad \text{dla przedziału ufności 95\%.}$$

W przypadku rzadkich chorób, dla których proporcja w populacji ogólnej jest niska, różnice w częstości występowania choroby na początku badania i przy jego końcu są minimalne – wtedy trzy miary, takie jak:
ryzyko zapadalności (*incidence risk*), częstość choroby (tempo zapadalności, *incidence rate*) i szansa wystąpienia choroby (*odds ratio, odds of disease*) – są jednakowe.

W przypadku rzadko spotykanej choroby:

$$OR = \frac{ad}{bc} = RR ,$$

natomiast w przypadku powszechnie spotykanej choroby:

$$OR = \frac{ad}{bc} \qquad 0 < |RR| < |OR|$$

W przypadku chorób powszechnie występujących te trzy miary są różne; chociaż przeważa tendencja, aby stosować współczynnik zapadalności jako miarę częstości choroby (*incidence rate*), wydaje się, że właściwsze byłoby zastosowanie ryzyka zapadalności (*incidence risk*), zwłaszcza jeżeli oceniamy protekcyjny efekt czynnika, tak jak np. w przypadku badania skuteczności szczepionek, kiedy mamy do czynienia z rozkładem odpowiedzi na zasadzie „wszystko albo nic": albo pełna protekcja u niektórych osób, albo żadna u pozostałych.

Ryzyko względne (jako ryzyko zapadalności) $RR = \frac{a/e}{c/f}$.

Ryzyko względne (jako częstość choroby) $RR = \frac{a/(osobo-lata\ działania\ czynnika)}{c/(osobo-lata\ niedziałania\ czynnika)}$.

Iloraz szans $OR = \frac{a/b}{c/d}$.

Ta nierównoważność wynika w głównej mierze ze schematu doboru grupy kontrolnej. Najczęściej stosowana jest procedura losowania kontroli spośród osób, które są zdrowe na etapie zakończenia badania. Jeżeli wylosowane w czasie trwania badania kontrole zachorują zanim badanie się zakończy, są one przemieszczane do grupy badanej. W takim wypadku wynikiem krzyżowego mnożenia komórek tabeli czteropolowej będzie zawsze iloraz szans, który jest większy niż ryzyko względne.

Występowanie zmiennych uwikłanych (współtowarzyszących)

Iloraz szans stosujemy przede wszystkim w badaniach kliniczno-kontrolnych (*case-control*). W planowaniu badań bardzo pożądane jest upewnienie się co do nieobecności dodatkowych, współtowarzyszących zmiennych, zwanych zmiennymi wikłającymi (uwikłanymi, *confounding variables*). Często nie jest możliwe wyeliminowanie niektórych zmiennych, nawet w sytuacjach jeżeli mamy pewność, że wpływają jako dodatkowy czynnik na wielkość badanej zależności. Takimi zmiennymi są na przykład wiek lub płeć. Byłoby

bezzasadne wyeliminowanie ich wpływu oraz świadome ignorowanie potencjalnego znaczenia struktury wiekowej czy rozkładu płci w badanych zależnościach. W sytuacjach, kiedy występują takie dodatkowe „wikłające" zmienne, można zastosować procedurę stratyfikacji (rozdziału na grupy) oraz policzyć interesującą nas zależność oddzielnie dla każdej warstwy/grupy (*stratum*).

Wydzielając kilka grup za względu na kategorie zmiennej „wikłającej" tworzymy warstwy (klasy) w jednej większej grupie i obliczamy interesujące nas statystyki osobno dla każdej wydzielonej klasy. Na podstawie tych „cząstkowych" wartości testu chi^2 obliczamy dopiero wartość ogólną, wspólną dla wszystkich warstw, zwaną stratyfikacyjnym ilorazem szans lub stratyfikacyjnym ryzykiem względnym (test Cochrana-Mantela-Haenszela):

ryzyko względne Mantela-Haenszela (przy stratyfikacji grup): $RR_{MH} = \dfrac{\sum_{i=1}^{k} \dfrac{a_i(c_i + d_i)}{N_i}}{\sum_{i=1}^{k} \dfrac{c_i(a_i + b_i)}{N_i}}$,

iloraz szans Mantela-Haenszela (przy stratyfikacji grup): $OR_{MH} = \dfrac{\sum(a_i d_i / N_i)}{\sum(b_i c_i / N_i)}$.

Wartość odpowiedniej statystyki chi^2 dla tak liczonych wartości ryzyka względnego czy ilorazu szans Mantela-Haenszela liczymy jako test chi^2 Cochrana-Mantela-Haenszela:

$$\chi_{CMH}{}^2 = \frac{\left[\sum \dfrac{a_i d_i - b_i c_i}{n_i}\right]^2}{\sum \dfrac{(a+b)_i (c+d)_i (a+c)_i (b+d)_i}{(n_i - 1)(n_i{}^2)}}$$

lub alternatywnie jako: $\chi_{CMH}{}^2 = \dfrac{\left[\sum \left(a_i - \dfrac{(a_i + b_i)(a_i + c_i)}{n_i} \right)\right]^2}{\sum \dfrac{(a+b)_i (c+d)_i (a+c)_i (b+d)_i}{(n_i - 1)(n_i{}^2)}}$

Jeżeli wzrasta liczba poziomów stratyfikacyjnych (k), to wygodniej jest policzyć wyraz stały w liczniku: $e_i = \dfrac{(a_i + b_i)(a_i + c_i)}{n_i}$

oraz w mianowniku: $v_i = \dfrac{(a+b)_i (c+d)_i (a+c)_i (b+d)_i}{(n_i - 1)(n_i{}^2)}$

Wtedy równanie ma postać: $\chi^2_{CMH} = \dfrac{\left[\sum (a_i - e_i)\right]^2}{\sum v_i}$

lub po uwzględnieniu poprawki Yatesa: $\chi^2_{CMH} = \frac{\left[\sum(a_i - e_i) - 0.5\right]^2}{\sum v_i}$

Niebezpieczeństwem procedury stratyfikacji jest utworzenie zbyt dużej liczby warstw, z których każda zawiera zbyt małe liczebności przypadków.

Alternatywnie, w sytuacjach gdy występują „nieusuwalne" wikłające zmienne, można zastosować modele skojarzone, tzn. takie gdzie dla każdego przypadku dobiera się odpowiadające kontrole, które pod względem wieku, płci, statusu genetycznego, czy jakiegokolwiek czynnika „wikłającego", są równocenne z grupą badaną.

Takie skojarzenie (pod względem zmiennych uwikłanych) przypadków i kontroli ma swój wyraz w procedurze analizy, a konkretnie w sposobie tabelaryzowania danych (obserwacji). Wartość ilorazu szans jest wtedy szacowana jako proporcja niezgodnych par przypadków i kontroli (zobacz przykłady w *„Części II – Uzupełnienia, przykłady i zadania"*):

$$\text{OR} = \text{iloraz niezgodnych par} = \frac{liczba\ par: przypadki_{narażone} - kontrole_{nienarażone}}{liczba\ par: przypadki_{nienarażone} - kontrole_{narażone}}$$

Jeżeli występuje kilka kontroli zamiast jednej skojarzonej z określonym przypadkiem, mają zastosowanie szczególne procedury oparte na statystyce testu chi^2 Mantela-Haenszela.

Gdy analizujemy kilka różnych czynników ryzyka i/lub kiedy występują dodatkowe zmienne uwikłane (współtowarzyszące) nierównocenne w badanym modelu, możemy zastosować metody warunkowej regresji logistycznej.

Podsumowanie metod opartych na statystyce testu chi^2 wykorzystywanych w badaniach populacyjnych kliniczno-kontrolnych przedstawia Tabela 7.

Kliniczne badania skuteczności szczepionek *(vaccine trials)*

Szczególnym rodzajem badań klinicznych są badania dotyczące skuteczności szczepionek. Idea tych badań jest trochę odmienna od niektórych badań klinicznych nad skutecznością leków: w badaniu uczestniczą wyłącznie zdrowi ochotnicy, z których część dostaje szczepionkę, a część nie. Skuteczność działania szczepionki określa się jako:

$$\text{Skuteczność szczepionki} = 1 - \frac{częstość\ zachorowań\ zaszczepionych}{częstość\ zachorowań\ niezaszczepionych} = 1 - \frac{1}{R}$$

gdzie RR oznacza ryzyko względne.
Przedział ufności obliczamy jako:

$$95\%CI = 1 - \frac{1}{RR^{1\pm 1.96/\chi}}$$

Zobacz też przykłady w *„Części II – Uzupełnienia, przykłady i zadania"*.

Tab. 7. Analiza statystyczna w badaniach kliniczno-kontrolnych.

schemat dobierania grupy kontrolnej	**pojedynczy czynnik ryzyka**	**wiele czynników ryzyka/ uwzględnianie zmiennych „wikłających"**
a) jedna kontrola do jednego przypadku		
przypadkowe	tabela 2 x 2: czynnik ryzyka *x* przypadek/kontrola standardowy test χ^2 OR = iloczyn krzyżowy, ad/bc	regresja logistyczna lub analiza stratyfikacyjna
dopasowanie sparowane	tabela 2 x 2 niezgodności par kontroli i przypadków w odniesieniu do czynnika ryzyka test χ^2 McNemara OR dla przeciwstawnych par = $= \frac{przypadki\ TAK, kontrole\ NIE}{przypadki\ NIE, kontrole\ TAK}$	warunkowa regresja logistyczna
dopasowanie stratyfikacyjne	analiza stratyfikacyjna, tabela 2 x 2 dla każdej warstwy test χ^2 Mantela-Haenszela, $OR_{MH} = \frac{\sum(ad/n)}{\sum(bc/n)}$	regresja logistyczna lub analiza stratyfikacyjna
b) wiele kontroli do jednego przypadku		
przypadkowe	jak wyżej	jak wyżej
dopasowane ściśle przypadki do kontroli	specjalne warianty testu χ^2 Mantela-Haenszela	warunkowa regresja logistyczna
dopasowanie stratyfikacyjne	jak wyżej	jak wyżej
c) przedziały ufności		
wszystkie schematy	95% = $OR^{(1\pm1.96/\chi)}$	obliczyć na podstawie współczynników regresji

Czułość, swoistość, wartości predykcyjne

Charakteryzując jakąś procedurę czy metodę analityczną staramy się intuicyjnie wykazać jej przewagę nad innymi alternatywnymi metodami oraz ocenić trafność diagnostyczną danej metody. Alternatywnie, wiedząc, że bardzo wiarygodne specjalistyczne nowoczesne badania (które uznajemy jako badania wzorcowe, referencyjne) są niezwykle kosztowne, staramy się często opracować i z powodzeniem wykorzystywać prostsze i tańsze metody diagnostyczne. Zależy nam przede wszystkim na określeniu:

- jak czuła jest dana metoda, to znaczy jaka jest szansa, że posługując się tą metodą uzyskamy taki wynik, jakiego się spodziewamy,
- jak selektywna czy swoista jest metoda, to znaczy jaka jest szansa, że uzyskamy wynik nie-pozytywny zawsze wtedy, gdy nie spodziewamy się go uzyskać.

Zdefiniowanie tych dwóch miar wskazuje na problem przekłamywania rzeczywistości przez wynik danej metody diagnostycznej/analitycznej. Staramy się zawsze maksymalnie zwiększyć wybiórczość metody (swoistość, *specificity*) i maksymalnie podnieść jej wrażli-

wość (czułość, *sensitivity*). Zawsze jednak musimy liczyć się z ryzykiem uzyskiwania z pewną częstością wyników fałszywie dodatnich (*false positive*) (w sytuacji gdy wynik będzie pozytywny nawet wtedy gdy takiego nie oczekujemy) oraz wyników fałszywie ujemnych (*false negative*) (wtedy gdy metoda „przeoczy" występowanie wyniku dodatniego).

		rzeczywiste występowanie badanego czynnika	
		występuje	***nie występuje***
wynik testu	***dodatni***	czułość	fałszywie dodatnie
	ujemny	fałszywie ujemne	swoistość

Test czuły to taki, który rzadko pomija w badaniu osoby naprawdę chore, a test swoisty rzadko kwalifikuje osoby naprawdę zdrowe jako chore. Test czuły wybierzemy wtedy, gdy konsekwencje niewykrycia choroby są bardzo poważne, a także w przypadku badań skriningowych, gdy prawdopodobieństwo choroby jest małe, czyli staramy się wykryć przypadki choroby u osób bez wyraźnych objawów. Testem swoistym posłużymy się w celu potwierdzenia naszego wstępnego rozpoznania (diagnozy), szczególnie w przypadkach, gdy mylne zakwalifikowanie osoby jako przypadku fałszywie dodatniego niesie wymierne szkody natury finansowej, psychicznej, itp.

Miary charakteryzujące trafność diagnostyczną metody obliczamy na podstawie liczności w komórkach tablicy czteropolowej:

		świat realny (np. wynik badania wzorcowego)		
		obecny	***nieobecny***	
wynik testu	***obecny***	*a*	*b*	***a+b***
	nieobecny	*c*	*d*	***c+d***
		a+c	***b+d***	

Do najczęściej stosowanych w praktyce miar zaliczamy:
- czułość testu (*test sensitivity*),
- swoistość testu (*test specificity*),
- dokładność rozpoznania (*diagnostic precision*),
- rzeczywistą częstość choroby (*true frequency of disease*),
- przewidywaną częstość choroby z predykcją dodatnią lub ujemną (*predicted frequency of disease with positive/negative prediction*),
- prawdopodobieństwo wyniku dodatniego (*likelihood of positive result*),
- prawdopodobieństwo wyniku ujemnego (*likelihood of negative result*),
- wartość predykcji wyników ujemnych (*prediction of negative results*) oraz,
- wartość predykcji wyników dodatnich (*prediction of positive results*).

Ich wartości możemy wyrażać albo jako proporcje albo w procentach:

$$\textit{czułość} = \frac{a}{a+c} \qquad \textit{swoistość} = \frac{d}{b+d}$$

$$\textit{dokładność rozpoznania} = \frac{a+d}{a+b+c+d}$$

$$\textit{rzeczywista częstość choroby} = \frac{a+c}{a+b+c+d}$$

$$\textit{przewidywana częstość choroby u osób z predykcją dodatnią} = \frac{a}{a+b}$$

$$\textit{przewidywana częstość choroby u osób z predykcją ujemną} = \frac{c}{c+d}$$

$$\textit{szansa wyniku fałszywie dodatniego} = \frac{b}{b+d}$$

$$\textit{szansa wyniku fałszywie ujemnego} = \frac{c}{a+c}.$$

Jeżeli uwzględnimy liczbę przypadków danej choroby występujących w danym momencie w populacji, to możemy obliczyć przewidywane wartości prawdopodobieństw pojawienia się wyniku ujemnego lub wyniku dodatniego w badaniu przeprowadzonym w próbie z takiej populacji. Prawdopodobieństwo, że występuje choroba w przypadku, gdy wynik testu jest ujemny (przewidywana wartość ujemna, wartość predykcji wyników ujemnych, PWU) wyliczymy jako:

$$PWU = \frac{[(1 - \textit{czułość}) \times \textit{chorobowość}]}{[(1 - \textit{czułość}) \times \textit{chorobowość}] + [\textit{czułość} \times (1 - \textit{chorobowość})]}$$

Z kolei, prawdopodobieństwo, że występuje choroba w przypadku, gdy wynik testu jest dodatni (przewidywana wartość dodatnia, wartość predykcji wyników dodatnich, PWD) wyliczymy jako:

$$PWD = \frac{[\textit{czułość} \times \textit{chorobowość}]}{[\textit{czułość} \times \textit{chorobowość}] + [(1 - \textit{czułość}) \times (1 - \textit{chorobowość})]}$$

Niestety w praktyce diagnostycznej często okazuje się, że testy bardzo czułe charakteryzują się niewystarczającą swoistością, i odwrotnie, a znalezienie zadowalającego kompromisu jest trudne i wymaga dużego doświadczenia. Jedną z metod, które mogą się okazać pomocne przy takich rozstrzygnięciach jest graficzna metoda krzywych typu ROC (*receiver operating characteristic curve*). Wiedza na temat analizy krzywych ROC ewoluuje od kilkudziesięciu lat i omówienie zastosowania tej techniki do różnych konkretnych przypadków w praktyce przekracza ramy niniejszego opracowania (zobacz też „*Literatura uzupełniająca*").

W celu bliższego wstępnego zapoznania się z tą metodą Czytelnik powinien także przeanalizować przykłady w „*Części II – Uzupełnienia, przykłady i zadania*".

Podsumowanie

- Współczynnik ryzyka względnego (RR) to najlepsza miara siły związku między zadziałaniem czynnika (chorobotwórczego) a wystąpieniem zmiany (chorobowej). Czym wyższa jego wartość, tym większe prawdopodobieństwo, że zależność taka jest przyczynowo-skutkowa. Ryzyko przypisane (AR) daje z kolei najlepsze pojęcie o wzroście ryzyka zapadalności po wystawieniu na działanie czynnika.
- Współczynnik ryzyka względnego (RR) wykorzystujemy najczęściej w badaniach prospektywnych, kiedy mamy możliwość manipulowania czynnikiem sprawczym zmian.

- Wartość i sposób szacowania RR zależy od naszego modelu badawczego oraz od tego czy badania dotyczą rzadkiej choroby czy choroby powszechnie występującej w populacji objętej badaniami.
- Miarą zależności w badaniach kliniczno-kontrolnych (z reguły retrospektywnych) jest iloraz szans (OR), który wyraża, ile razy szansa zachorowania wśród osób poddanych działaniu czynnika jest wyższa od szansy zachorowania wśród osób nie poddanych działaniu czynnika. Wysoka wartość OR jest wskazaniem, że badany czynnik jest silnym czynnikiem ryzyka wystąpienia choroby.
- OR może być także użytecznym wskaźnikiem przy wyrażaniu, ile razy jakaś patologiczna cecha/zmiana występuje częściej u osób chorych w porównaniu z grupą kontrolną.
- Wskaźniki takie jak czułość, swoistość czy wartość predykcyjna wyniku stanowią podstawowe narzędzie analizy i oceny jakości danych diagnostycznych.
- Metody stosowane w badaniach populacyjnych oparte są na statystyce testu chi^2 i dlatego podlegają wszystkim założeniom i ograniczeniom typowym dla tego testu.

Część II

Uzupełnienia, przykłady i zadania

Zadania, przykłady i uzupełnienia zawarte w tej części, jak również arkusze danych, do których odwołują się wybrane przykłady, zostały także umieszczone na płycie CD załączonej do niniejszego opracowania.

„Błędne przeświadczenie, że nauka prowadzi koniec końców do ostatecznych wyjaśnień, wywołuje u naukowców przekonanie, że popełniają poważne wykroczenie, ogłaszając hipotezy, które okażą się w końcu fałszywe."

Karl Raimund Popper

Rozdział 14

Statystyki podstawowe, rachunek prawdopodobieństwa i rozkłady zmiennych

„Wszystkie organizmy są stale, dniami i nocami, zaabsorbowane rozwiązywaniem problemów ..."

Karl Raimund Popper

Miary położenia i rozproszenia

Przykład 1

Należy policzyć miary centralne, odchylenie standardowe oraz współczynnik zmienności dla wartości pomiarów stężenia całkowitego cholesterolu w osoczu krwi (mg/100 ml) u zdrowych ochotników:

221	272	192	240	305	264	199	211
256	243	317	187	211	195	292	228

Ponieważ suma wszystkich wartości $\Sigma x = 3833$, średnia arytmetyczna, $\sum x/n = 3833/16 = 239.6$.

Aby policzyć średnią geometryczną obliczamy wartości $u_i = \log(x_i)$:

2.27	2.28	2.29	2.30	2.32	2.32	2.34	2.36
2.38	2.39	2.41	2.42	2.43	2.47	2.48	2.50

$\bar{u} = \frac{\sum_{i=1}^{n} u_i}{n} = 37.98/16 = 2.37$, oraz $\bar{x}_G = antilog(\bar{u}) = 10^{\bar{u}} = 10^{2.37} = 234.4$

W celu obliczenia mediany dane sortujemy w porządku rosnącym:

187	192	195	199	211	211	221	228
240	243	256	264	272	292	305	317

Me = (*n*+1)/2 = 17/2 = $8\frac{1}{2}$ wartość w szeregu = średnia 8 i 9, wartości = (228+240)/2 = 234.
Modalna nie istnieje dla tej grupy danych, gdyż wszystkie wartości są nieidentyczne.
Odchylenie standardowe wynosi:

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{(n-1)}} = \sqrt{\frac{25725.94}{16-1}} = 41.4$$

lub liczone w inny sposób:

$$s = \sqrt{\frac{\sum x^2 - (\sum x)^2 / n}{(n-1)}} \sqrt{\frac{943969 - (3833)^2/16}{16-1}} = \sqrt{\frac{943969 - 918243.06}{15}} = 41.4$$

i odpowiednio wariancja s^2 = 1715.0625.
Współczynnik zmienności wynosi: $CV = \frac{s}{x} \times 100\%$ = 41.4/239.6 x 100% = 17.3%

Przykład 2

U 10 dzieci zbadano stężenia przeciwciał w osoczu po miesiącu przeprowadzenia u nich szczepienia na odrę. W tabeli poniżej przedstawiono miana roztworów jako odwrotności kolejnych rozcieńczeń roztworów użytych do miareczkowania. Jakie jest średnie miano przeciwciał w osoczu w badanej grupie dzieci?

dziecko	miano przeciwciał		rozcieńczenie nr
1	8	2^3	3
2	16	2^4	4
3	16	2^4	4
4	32	2^5	5
5	8	2^3	3
6	128	2^7	7
7	16	2^4	4
8	32	2^5	5
9	32	2^5	5
10	16	2^4	4

u = nr rozcieńczenia = $\log_2$ (miano przeciwciał)
średnia geometryczna miana przeciwciał = $2^{\text{średni nr rozcieńczenia}}$
$\bar{u}$ = 4.4, zatem $\bar{x}_G = 2^{4.4}$ = 21.1

Przykład 3

W doświadczeniu badano wpływ metali ciężkich w diecie na przeżywalność szczurów. Obserwacje prowadzono w grupie 20 zwierząt w ciągu 30 dni. Uzyskano następujące wyniki:

12	16	22	28	18	25	21	8
24	14	27	30	22	20	11	17

4 zwierzęta pozostały żywe do końca prowadzenia obserwacji. Ile dni wynosiła średnio przeżywalność zwierząt?

Do obliczeń wykorzystujemy średnią harmoniczną:

$$\overline{X}_H = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

$\sum_{i}^{n} \frac{1}{x_i}$ = 1/12 + 1/16 + 1/22 + 1/28 + 1/18 + 1/25 + 1/21 + 1/8 + 1/24 + 1/14 + 1/27+ 1/30 + 1/22 + 1/20 + 1/11 + 1/17 +1/∞ +1/∞ +1/∞ +1/∞ = 0.0833 +.0625 + 0.0455 + 0.0357 + 0.0556 + 0.04 + 0.0476 + 0.125 + 0.0417 + 0.714 + 0.037 + 0.0333 + 0.0454 + 0.05 + 0.091 + 0.0588 + 0 + 0 + 0 + 0 = 0.259

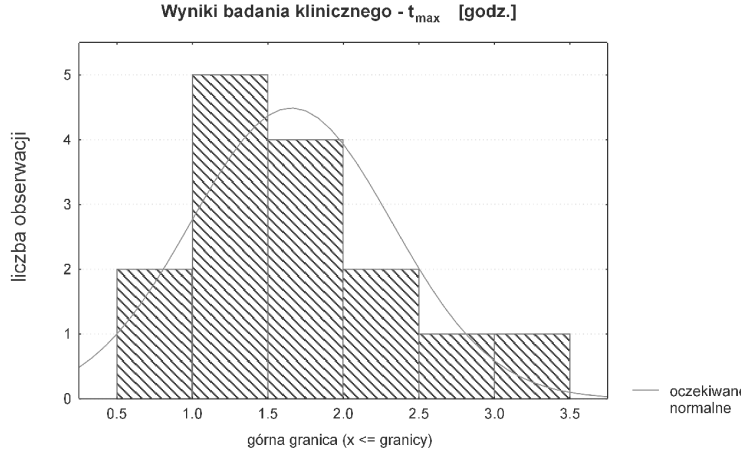
$$\overline{X}_H = \frac{20}{0.259} = 77 \text{ dni}$$

Przykład 4

W tabeli przedstawiono czasy, dla których stężenie leku w osoczu krwi przyjmowało maksymalną wartość. Należy obliczyć medianę, średnią arytmetyczną i średnią geometryczną oraz odpowiedzieć na pytanie czy mamy prawo sądzić, że rozkład wyników jest prawoskośny?

wyniki badania klinicznego – t_{max} (godz.)					
pacjent	t_{max}	pacjent	t_{max}	pacjent	t_{max}
1	1.41	6	1.96	11	1.62
2	1.81	7	0.78	12	1.15
3	3.25	8	1.51	13	2.03
4	1.37	9	1.18	14	2.21
5	1.09	10	2.56	15	0.91

Wykres rozkładu danych przedstawia się następująco:



Obliczamy miary centralne:
średnia arytmetyczna, $\sum x / n = 24.84/15 = 1.66$
obliczamy wartości $u_i = \log(x_i)$ i na ich podstawie średnią geometryczną:

–0.108	–0.041	0.037	0.061	0.072	0.137	0.149	0.179
0.210	0.258	0.292	0.307	0.344	0.408	0.512	

$$\bar{u} = \frac{\sum_{i=1}^{n} u_i}{n} = 2.818/15 = 0.188, \text{ oraz } \bar{x}_G = anti\log(\bar{u}) = 10^{\bar{u}} = 10^{0.188} = 1.54$$

W celu obliczenia mediany dane sortujemy w porządku rosnącym:

0.78	0.91	1.09	1.15	1.18	1.37	1.41	**1.51**
1.62	1.81	1.96	2.03	2.21	2.56	3.25	

$Me = (n+1)/2 = 16/2 = 8$, wartość w szeregu = 1.51

Na podstawie wykresu oraz wartości miar centralnych (średnia arytmetyczna jest zdecydowanie wyższa niż średnia geometryczna, która jest bliższa medianie) mamy prawo sądzić, że rozkład wyników jest prawoskośny.

Przykład 5

W tabeli przedstawiono wyniki objętości płytek krwi uzyskane w losowo wybranych próbach o liczebnościach 1-5. Należy policzyć średnią wartość objętości płytek krwi w grupie zbadanych osób.

x_i [fL]	f_i	$f_i x_i$
7.0	1.0	7.0
7.2	0.0	0.0
7.4	3.0	22.2
7.6	1.0	7.6
7.8	5.0	39.0
8.0	2.0	16.0
8.2	11.0	90.2
8.4	6.0	50.4
8.6	2.0	17.2
8.8	9.0	79.2
9.0	7.0	63.0
9.2	2.0	18.4
9.4	8.0	75.2
9.6	0.0	0.0
9.8	4.0	39.2
	$\Sigma f_i = 61$	$\Sigma f_i x_i = 524.6$

$k = 15$ grup $\quad \Sigma f_i = 61$ $\quad \bar{x} = \frac{\sum_{i=1}^{k} f_i x_i}{n} = \frac{524.6}{61} = 8.6$

Rachunek prawdopodobieństwa

Przykład 6

Jeżeli w pewnym regionie prawdopodobieństwo wystąpienia choroby A wynosi 55%, a prawdopodobieństwo wystąpienia choroby B 62%, to jakie jest prawdopodobieństwo wystąpienia u przypadkowo wylosowanej osoby jednej z chorób (albo choroby A albo choroby B)?

Z oczywistych względów, prawdopodobieństwo wystąpienia choroby A albo choroby B nie jest prostą suma prawdopodobieństw, 0.55 + 0.62 = 1.17, ponieważ łączne prawdopodobieństwo nie może mieć wartości wyższej niż jeden, a poza tym nie możemy liczyć prawdopodobieństwa wystąpienia zarówno choroby A, jak i choroby B, podwójnie. Zapiszemy zatem:

$$P \text{ (A lub B lub zarówno A jak i B)} = P \text{ (A)} + P \text{ (B)} - P \text{ (zarówno A jak i B)}$$

Jeżeli dwie choroby występują łącznie, to prawdopodobieństwo takiego zdarzenia wynosi:

$$P \text{ (zarówno A jak i B)} = 0.55 \text{ x } 0.62 = 0.341$$

Zatem całkowite

$$P \text{ (A lub B lub zarówno A jak i B)} = 0.55 + 0.62 - 0.341 = 0.829$$

Przykład 7

Zakładając, że jesteś nosicielem pewnego rzadkiego allelu, jakie jest prawdopodobieństwo, że któreś z twoich prawnucząt będzie nosicielem tego samego rzadkiego allelu pochodzącego od ciebie?

Prawdopodobieństwo, że jakiekolwiek dziecko otrzyma od jednego ze swoich rodziców jakiś rzadki allel (a więc taki, który występuje na pewno jako pojedyncza kopia w genotypie) wynosi 1/2. Ponieważ trzy pokolenia dzielą ciebie i twoje prawnuczęta, to szansa, że któreś z nich otrzyma ten allel wynosi:

$$1/2 \times 1/2 \times 1/2 = 1/8$$

Przykład 8

Oszacowano dla populacji Murzynów amerykańskich, że jeden na 575 Murzynów ma anemię sierpowatą. W chorobie tej występują nieprawidłowe cząsteczki hemoglobiny S, co sprawia, że krwinki czerwone mają charakterystyczny sierpowaty kształt w warunkach niskiego ciśnienia parcjalnego tlenu. Choroby tej prawie wcale nie spotyka się u nie-Murzynów żyjących w USA. Okazało się, że chorzy są nosicielami dwóch recesywnych

genów HbS. Wiedząc, że w jednym z miast amerykańskich z populacją 64 tysiące mieszkańców co czwarty mieszkaniec jest Murzynem, jakie jest prawdopodobieństwo, że przypadkowo wylosowana osoba żyjąca w tym mieście będzie miała anemię sierpowatą oraz ilu mieszkańców tego miasta mogłoby mieć tą chorobę?

$$1/575 \times 1/4 = 1/2300$$

Prawdopodobieństwo takie wynosi 1/2300 = 0.04%. W 64-tysięcznym mieście, statystycznie biorąc, 64000/2300 ≈ 28 mieszkańców może chorować na anemię sierpowatą.

Przykład 9

W pewnym kraju około 10% ludzi jest nosicielami grupy krwi 0. Jaka jest szansa, że w losowo wybranym małżeństwie oboje małżonkowie są nosicielami grupy krwi 0?

$$1/10 \times 1/10 = 1/100$$

Szansa taka wynosi 1%.

Przykład 10

Jeżeli oboje małżonkowie są nosicielami grupy krwi B, to jakie jest prawdopodobieństwo, że ich pierwsze dziecko będzie miało grupę krwi 0, jeśli wiemy, że 3 na każdych 7 nosicieli grupy krwi B jest heterozygotami?

Podstawowym warunkiem urodzenia się u tej pary dziecka z grupą krwi 0 jest, aby oboje rodzice posiadali genotyp B0 układu AB0 grup krwi. Szansa tego zdarzenia wynosi $3/7 \times 3/7 = 9/49$. Urodzenie dziecka z grupą krwi 0 z rodziców, z których każde ma genotyp B0 wynosi $1/2 \times 1/2 = 1/4$. Zatem jednoczesne prawdopodobieństwo obu zdarzeń będzie wynosić:

$$9/49 \times 1/4 = 9/196$$

Przykład 11

Za występowanie żółtaczki hemolitycznej odpowiada obecność dominującego genu o penetracji około 10% (tzn. że choroba wystąpi jedynie u 1 na około 10 nosicieli tego genu). Jaka jest szansa, że dziecko heterozygotycznego mężczyzny pod względem tego genu oraz zdrowej homozygotycznej kobiety będzie nosicielem tej cechy?

Szansa uzyskania przez dziecko genu od ojca wynosi $1/2$, zaś szansa wystąpienia choroby u nosiciela genu $1/10$. Dlatego szansa, że dziecko będzie miało żółtaczkę hemolityczną wynosi:

$$1/2 \times 1/10 = 1/20$$

Przykład 12

Para planująca posiadanie czworga dzieci chciałaby mieć po dwa z każdej płci. Jaka jest szansa, że ich plany się spełnią?

Szansa posiadania dziecka określonej płci wynosi 1/2, czyli prawdopodobieństwo urodzenia chłopca wynosi $a = 1/2$ oraz prawdopodobieństwo urodzenia dziewczynki też wynosi $b = 1/2$. Jeżeli założymy, że podczas każdego porodu będzie im się rodziło tylko jedno dziecko, to prawdopodobieństwo różnych wariantów kombinacji płci dzieci urodzonych w n kolejnych porodach jest rozwinięciem dwumianu Newtona:

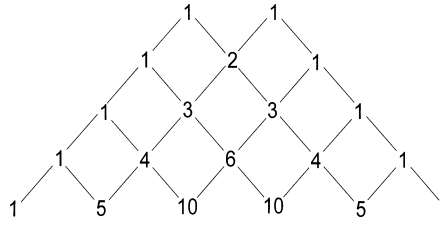
$$(a+b)^n$$

Zgodnie ze schematem trójkąta Pascala (Ryc. 1), dla czterech urodzin rozwinięciem tego dwumianu jest równanie:

$$(a+b)^4 = a^4 + 4a^3b + 6a^2b^2 + 4ab^3 + b^4$$

Interesuje nas kombinacja: dwie dziewczynki oraz dwóch chłopców, zatem z rozwinięcia dwumianu wybieramy wyraz $6a^2b^2$.

Prawdopodobieństwo, że małżeństwo planujące posiadanie czworga dzieci będzie miało po dwa z każdej płci wynosi $6(0.5)^2(0.5)^2 = 0.375$ czyli 37.5%.



Ryc. 12. Trójkąt Pascala jest prostą metodą znajdowania współczynników rozwinięcia dwumianu Newtona $(a+b)^n$. W każdym rzędzie druga liczba oznacza potęgę dwumianu, n. Każdy współczynnik jest sumą liczb leżących w rzędzie ponad nim. Alternatywnie, współczynniki możemy także obliczyć w inny sposób: współczynnik pierwszego wyrazu będzie zawsze równy 1, współczynnik drugiego wyrazu jest zawsze równy wykładnikowi pierwszego wyrazu, współczynniki następnych wyrazów liczymy jako: (współczynnik poprzedzającego wyrazu x wykładnik potęgi dla a w poprzedzającym wyrazie)/pozycja poprzedzającego wyrazu, np. dla rozwinięcia dwumianu z wykładnikiem 4 będziemy mieli: 1 – (1 x 4: 1 = 4) – (4 x 3: 2 = 6) – (6 x 2: 3 = 4) – (4 x 1: 4 = 1).

Przykład 13

Dwoje ludzi z normalną pigmentacją skóry posiada dwójkę dzieci albinotycznych i jedno normalne. Wiedząc, że szansa urodzenia albinosa z dwojga ludzi o normalnej pigmentacji skóry wynosi 1/4, jakie jest prawdopodobieństwo zaistnienia takiej kombinacji?

Oboje rodzice są oczywiście heterozygotami pod względem posiadania genu na albinizm, zatem szansa spłodzenia przez taką parę albinotycznego dziecka wynosi 1/4, zaś szansa posiadania nie-albinotycznego dziecka – 3/4.

Z rozwinięcia dwumianu:

$$(a+b)^3 = a^3 + 3a^2b + 3ab^2 + b^3$$

wybieramy wyraz $3a^2b$. Zatem interesujące nas prawdopodobieństwo wynosi:

$3 \times (1/4)^2 \times (3/4) = 3 \times (1/16) \times (3/4) = 9/64 = 0.140625$ (niecałe 15%).

Przykład 14

Pewien mężczyzna, który jest heterozygotą $Pl^{A1/A2}$ płytkowej glikoproteiny IIIa, posiada pięcioro dzieci z kobietą, która jest normalną homozygotą $Pl^{A1/A1}$. Troje z dzieci tej pary jest heterozygotami $Pl^{A1/A2}$ (tak jak ojciec), zaś dwoje homozygotami $Pl^{A1/A1}$ (tak jak matka). Jakie jest prawdopodobieństwo takiego zdarzenia?

Szansa urodzenia dziecka z genotypem $Pl^{A1/A2}$ z takiej pary rodziców wynosi:

Pl^{A2} (jedyna kopia od ojca) x Pl^{A1} (pierwsza kopia od matki)
lub Pl^{A2} (jedyna kopia od ojca) x Pl^{A1} (druga kopia od matki),
czyli $a = (1/2 \times 1/2) + (1/2 \times 1/2) = 1/4 + 1/4 = 1/2$

Podobnie, szansa urodzenia dziecka z genotypem $Pl^{A1/A1}$ u takiej pary rodziców wynosi:

Pl^{A1} (jedyna kopia od ojca) x Pl^{A1} (pierwsza kopia od matki)
lub Pl^{A1} (jedyna kopia od ojca) x Pl^{A1} (druga kopia od matki),
czyli $b = (1/2 \times 1/2) + (1/2 \times 1/2) = 1/4 + 1/4 = 1/2$

Z rozwinięcia dwumianu:

$$(a+b)^5 = a^5 + 5a^4b + 10a^3b^2 + 10a^2b^3 + 5ab^4 + b^5$$

wybieramy wyraz $10a^3b^2$. Jego rozwiązaniem jest interesująca nas wartość prawdopodobieństwa, które wynosi:

$10 \times (1/2)^3 \times (1/2)^2 = 10 \times (1/8) \times (1/4) = 10/32 = 0.3125$ (ponad 30%).

Przykład 15

Załóżmy, że nowotwory przyczyniają się do śmierci w 10% przypadków. W historii rodziny obejmującej 64 zmarłe osoby, 13 zmarło na raka. Czy statystyka śmiertelności z powodu nowotworu w tej rodzinie przystaje do statystyki populacji ogólnej?

całkowita liczebność (n) 64 osoby	
oczekiwana śmiertelność na nowotwór (p)	10% czyli 6.4
– nie na nowotwór (q)	64 – 6.4 = 57.6
obserwowana śmiertelność	
z powodu nowotworu	13
nie z powodu nowotworu	51

różnica między obserwowaną i oczekiwaną (d) $13 - 6.4 = 6.6$

błąd standardowy różnicy $SE = \sqrt{\frac{p*q}{n}} = \sqrt{\frac{6.4*57.6}{64}} = 2.4$

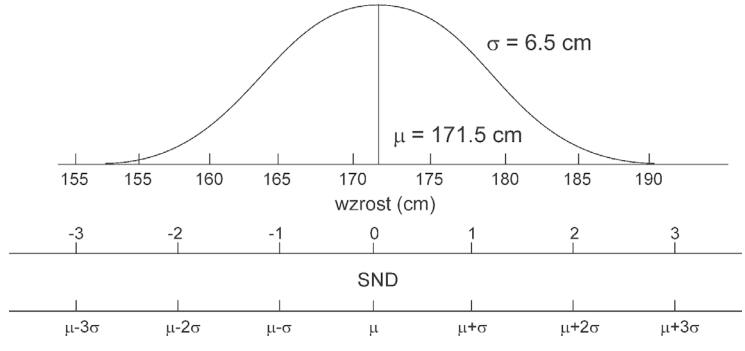
wartość statystyki testu t dla proporcji $t = \frac{d}{SE} = \frac{6.6}{2.4} = 2.75$

Ponieważ obliczone $t = 2.75 > t_{0.05} = 1.96$, zatem możemy odrzucić hipotezę, według której śmiertelność z powodu nowotworów przystaje do statystyki populacji ogólnej.

Rozkład normalny

Przykład 16

Rozkład wzrostu w grupie studentów Akademii Wychowania Fizycznego, posiadający charakterystykę rozkładu normalnego, przedstawiono na Ryc. 13.



Ryc. 13. Rozkład wzrostu w badanej populacji studentów oraz interpretacja wartości z standaryzowanego rozkładu normalnego (SND).

Na podstawie tego przykładu zapoznamy się z logiką obliczeń dotyczących charakterystyki rozkładu normalnego.
Wiedząc, że:

$$\text{SND} = (\text{wzrost} - 171.5)/6.5,$$

możemy zapisać, że dla naszego przykładu:

$$\text{wzrost} = 171.5 + (6.5 \times \text{SND})$$

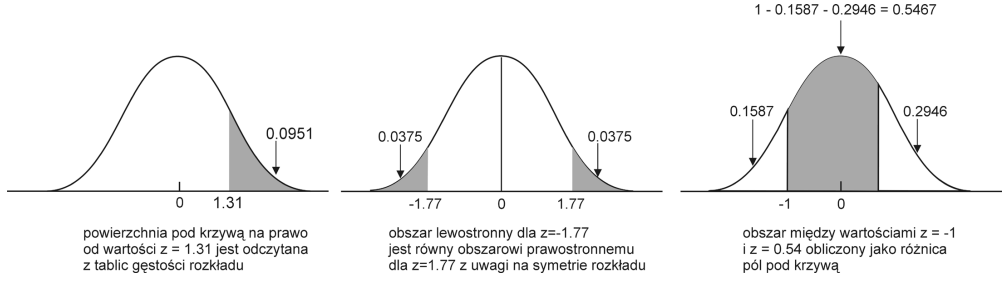
Znając wartości parametrów rozkładu normalnego: średniej i odchylenia standardowego, możemy łatwo oszacować wartość zmiennej (wzrostu) w dowolnym miejscu pod krzywą. Praktyczną zaletą standaryzacji rozkładu normalnego jest to, że do charakterystyki dowolnej zmiennej o rozkładzie normalnym – z określoną wartością średniej i SD – potrzebujemy jedynie tablic rozkładu znormalizowanego, a nie wszystkich możliwych

kombinacji różnych wartości średnich i SD. Dwa najczęściej spotykane rodzaje tablic to tablice dystrybuanty rozkładu (powierzchni pod krzywą dla określonej wartości zmiennej) oraz tzw. wartości krytycznych rozkładu (punktów frakcji procentowych pola pod krzywą).

Tablice dystrybuanty rozkładu normalnego są stosowane przy określaniu proporcji (frakcji populacji), dla których zmienna przyjmuje wartości z określonego zakresu. Na przykład, chcemy określić jaka część populacji studentów posiada wzrost większy od 180 cm. Dla takich obserwacji wartość z standaryzowanego rozkładu normalnego wynosi:

$$z = \frac{180 - 171.5}{6.5} = 1.31$$

Oznacza to, że do rozważanej podgrupy będą należeli studenci charakteryzujący się wzrostem większym od średniego dla grupy o wartość przynajmniej 1.31 odchyleń standardowych. Z tablic dystrybuanty odczytujemy, że powierzchnia pod krzywą rozkładu normalnego na prawo od 1.31 (tzn. od 1.31 do +∞) wynosi 0.0951 czyli 9.51% całości (Ryc. 14).



Ryc. 14. Obliczanie proporcji populacji na podstawie dystrybuanty rozkładu normalnego.

Analogicznie, proporcja populacji studentów niższych od 160 cm równa się wartości dystrybuanty w punkcie:

$$z = \frac{160 - 171.5}{6.5} = -1.77$$

która wynosi 0.0375 czyli 3.75% (Ryc. 14). Oznacza to, że u 9.51% wszystkich studentów możemy oczekiwać wzrostu większego od 180 cm, a u 3.75% całości – wzrostu mniejszego od 160 cm.

W podobny sposób możemy ocenić proporcje populacji studentów, których wzrost mieści się w zakresie od 165 cm do 175 cm (Ryc. 14). Odpowiednie wartości z i proporcje pola pod krzywą (dystrybuanty) wynoszą:

$$z = \frac{165 - 171.5}{6.5} = -1$$

Według tablic dystrybuanty proporcja studentów ze wzrostem poniżej 165 cm wynosi 0.1587.

$$z = \frac{175 - 171.5}{6.5} = 0.54$$

Według tablic dystrybuanty proporcja studentów ze wzrostem powyżej 175 cm wynosi 0.2946.

Proporcja studentów ze wzrostem w granicach 165-175 cm = 1 – proporcja osób ze wzrostem poniżej 165 cm – proporcja osób ze wzrostem powyżej 175 cm = 1 – 0.1587 – 0.2946 = 0.5467 czyli 54.67%.

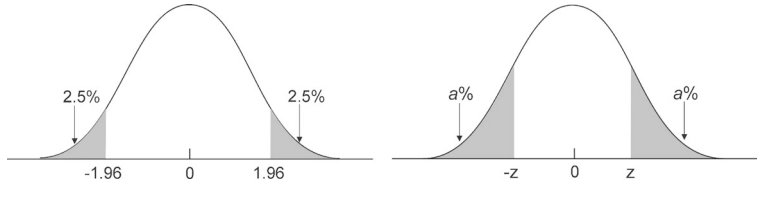
Na podstawie tablic dystrybuanty możemy także obliczyć jaka jest wartość zmiennej (wzrostu), która przekracza 5% (0.05) wszystkich (najwyższych) studentów. W tablicach odnajdujemy, że wartości prawdopodobieństwa 0.05 odpowiada wartość $z = 1.64$. Zatem ze wzoru:

$$x = \mu + z \times \sigma$$

obliczamy, że 5% wszystkich studentów posiada wzrost równy lub przekraczający 171.5 + 1.64 x 6.5 = 182.2 cm.

W przypadkach, gdy chcemy znać proporcje populacji charakteryzujące się wartościami zmiennych w określonym przedziale, wygodnie jest wyrażać wartość zmiennej jako wielokrotność SD, która dzieli ją od wartości średniej (Ryc. 13). Na przykład, $z = –1$ i $z = 1$ odpowiadają wartości mniejszej lub większej od średniej o jedną wartość odchylenia standardowego. Z uwagi na symetryczność rozkładu pole obszaru krytycznego dla $z = –1$ oraz $z = 1$ jest jednakowe i wynosi 0.1587. Wynika stąd, że 31.74% (2 x 15.87%) pola pod krzywą przypada na wartości oddalone od średniej o więcej niż 1 SD. Odpowiednio, 68.26% wszystkich wartości leży w zakresie średniej ± 1 SD. Odpowiednio, 4.55% rozkładu leży poza obszarem zakresu średnia ± 2 SD, zaś wewnątrz tego obszaru znajduje się 95.45% wyników.

Z tablic dystrybuanty możemy odczytać, że wartość z odpowiadająca 95% pola pod krzywą rozkładu normalnego (między $–z$ i z) wynosi 1.96 (Ryc. 15).



Ryc. 15. Dystrybuanta i proporcje rozkładu normalnego.

Odpowiednio, 5% wszystkich obserwacji (2.5% po każdej stronie) leży poza obszarem obejmującym zakres średniej ± 1.96xSD. Podobnie, 2.33 jest wartością dystrybuanty jednostronnej, a 2.58 obustronnej dla 1% obserwacji.

W tabeli na następnej stronie podano wartości dystrybuant dla wybranych prawdopodobieństw (proporcji) rozkładu normalnego.

proporcja pola pod krzywą	dystrybuanta jednostronna	dystrybuanta obustronna
50%	0.00	0.67
40%	0.25	0.84
30%	0.52	1.04
20%	0.84	1.28
10%	1.28	1.64
5%	1.64	1.96
2%	2.05	2.33
1%	2.33	2.58
0.5%	2.58	2.81
0.2%	2.88	3.09
0.1%	3.09	3.29
0.01%	3.72	3.89

Badanie normalności rozkładu

Przykłady procedur badania normalności rozkładu zostały uwzględnione w części *„Badanie dopasowania rozkładu"*.

Przedział ufności oraz wnioskowanie o średniej populacji ogólnej

Przykład 17

Planujemy poddać opryskowi insektycydem pola uprawy ziemniaków. Powierzchnia 5892 terenów uprawnych ziemniaków waha się w granicach od 9800 m² do 16145 m². Jeden litr insektycydu wystarcza do opryskania 50 m² powierzchni. W celu obliczenia ilości niezbędnego preparatu wybrano losowo 100 terenów uprawnych o średniej powierzchni 12140 m² i wartości odchylenia std. 815 m². Ponieważ jest mało prawdopodobne, aby średnia próby losowej ($\bar{x}$) pokrywała się idealnie ze średnią dla całej zbiorowości obejmującej 5892 pól (μ), pragniemy obliczyć zakres zmienności dla średniej, aby na tej podstawie oszacować ilość insektycydu niezbędnego do opryskania wszystkich pól uprawnych.

Precyzję oszacowania średniej populacji obliczamy jako błąd standardowy:

$$s/\sqrt{n} = 815/\sqrt{100} = 81.5 \text{ m}^2.$$

Istnieje prawdopodobieństwo 95%, że średnia dla populacji 5892 pól różni się od średniej próby 12140 m² mniej niż $o/z' \times s/\sqrt{n} = 1.96 \times 815\sqrt{100} = 1.96 \times 81.5 = 159.74$ m². Przedział ufności dla próby wynosi zatem:

$$\bar{x} \pm 1.96 \times s/\sqrt{n} = 12140 \pm 160 = 11980 \text{ do } 12300 \text{ m}^2.$$

Do dalszych obliczeń użyjemy górnej granicy tego zakresu. Skoro 1 litr insektycydu wystarcza na opryskanie 50 m² gruntu, to na opryskanie 5892 pól o średniej powierzchni 12300 m² potrzebujemy:

$$5892 \text{ x } 12300/50 = 1449432 \text{ litrów}$$

Istnieje 95% prawdopodobieństwo, że w przedziale 11980-12300 mieści się rzeczywista średnia populacji i 5% szans, że mieści się poza tym zakresem. Zatem istnieje 2.5% ryzyko (1/2 x 5%), że oszacowana ilość insektycydu będzie za mała. Aby zmniejszyć to ryzyko możemy dokonać tej samej oceny z prawdopodobieństwem 99%, ryzykiem popełnienia błędu 1% i ryzykiem zaniżenia potrzebnej ilości insektycydu 0.5%. Ponieważ prawdopodobieństwo 99% odpowiada wyższej wartości z' (2.58), zakres przedziału ufności odpowiadający 1% ryzyku popełnienia błędu jest szerszy i wynosi:

$$\bar{x} \pm 2.58 \times s/\sqrt{n} = 12140 \pm 210 = 11930 \text{ do } 12350 \text{ m}^2.$$

Dla 99.9% prawdopodobieństwa i 0.1% ryzyku popełnienia błędu zakres będzie wynosił odpowiednio od 11872 m^2 do 12408 m^2.

Przykład 18

Sześciu pacjentom cierpiącym na bóle migrenowe podawano nowy lek uśmierzający, który łagodził ból przez okres: 2.2, 2.4, 4.9, 2.5, 3.7 i 4.3 godziny. Naszym zadaniem jest określenie jaki jest przedział czasowy skuteczności leku.

$$\bar{x} = 3.3 \text{ godziny}, s = 1.13 \text{ godziny}, s/\sqrt{n} = 0.46 \text{ godziny}, n = 6, d.f. = n - 1 = 5$$

Z prawdopodobieństwem 95% (oraz istotnością 5%) przy 5 stopniach swobody wartość krytyczna testu t wynosi 2.57, zaś przedział ufności:

$$3.3 \pm 2.57 \text{ x } 0.46 = 3.3 \pm 1.2 = \text{od } 2.1 \text{ do } 4.5 \text{ godziny}$$

Rozkład dwumianowy

Przykład 19

Kobieta i mężczyzna są heterozygotami pod względem anemii sierpowatej (Ss), czyli mają po jednym allelu warunkującym anemię (S) i po jednym allelu normalnym (s). Małżeństwo to ma czworo dzieci. Jakie jest prawdopodobieństwo, że żadne, jedno, dwoje, troje lub wszystkie czworo dzieci będzie homozygotami pod względem tego genu (*SS*)?

Dla każdego z dzieci prawdopodobieństwo wystąpienia homozygoty *SS* jest równe prawdopodobieństwu otrzymania po jednym allelu S od każdego z rodziców, czyli 0.5 x 0.5 = 0.25. Odpowiednio, prawdopodobieństwo nie posiadania obu alleli *SS* (czyli posiadania po jednym allelu lub posiadania obu alleli s, Ss lub ss) wynosi 1 – 0.25 = 0.75. Oznaczmy cechę wystąpienia homozygoty *SS* jako A, zaś cechę niewystąpienia homozygoty *SS* (genotyp Ss lub ss) jako B. Zgodnie z naszymi wcześniejszymi obliczeniami $p = 0.25$.

Prawdopodobieństwo, że żadne z dzieci nie jest A (*SS*) (czyli $r = 0$) wynosi 0.75 x 0.75 x 0.75 x 0.75 = 0.75^4 = 0.3164. Prawdopodobieństwo, że jedno z dzieci jest *SS* jest prawdopodobieństwem, że (pierwsze dziecko jest *SS*, drugie, trzecie i czwarte są nie-*SS*) lub

(drugie dziecko jest *SS*, zaś pierwsze, trzecie i czwarte są nie-*SS*) lub (trzecie dziecko jest *SS*, zaś pierwsze, drugie i czwarte są nie-*SS*) lub (czwarte dziecko jest *SS*, natomiast pierwsze, drugie, trzecie są nie-*SS*). Zauważmy, że prawdopodobieństwo, że którekolwiek dziecko jest *SS* w sytuacji gdy pozostałe trzy nie są *SS*, wynosi 0.25×0.75^3. Ponieważ warianty takiego zdarzenia są cztery i nie mogą one wystąpić razem, całkowite prawdopodobieństwo wystąpienia jednego dziecka z *SS* wśród pozostałych nie-*SS* wynosi $4 \times 0.25 \times 0.75^3 = 0.4219$, zgodnie z zasadą addytywności prawdopodobieństw.

W analogiczny sposób możemy obliczyć prawdopodobieństwa dla wszystkich innych kombinacji w przypadku wystąpienia dwojga, trojga lub czworga dzieci *SS*:

liczba dzieci		**liczba**	
z genotypem *SS*	**z genotypem nie-*SS***	**możliwych kombinacji**	**prawdopodobieństwo**
0	4	1	$1 \times 1 \times 0.75^4 = 0.3164$
1	3	4	$4 \times 0.25 \times 0.75^3 = 0.4219$
2	2	6	$6 \times 0.25^2 \times 0.75^2 = 0.2109$
3	1	4	$4 \times 0.25^3 \times 0.75 = 0.0469$
4	0	1	$1 \times 0.25^4 \times 1 = 0.0039$
			razem 1.0000

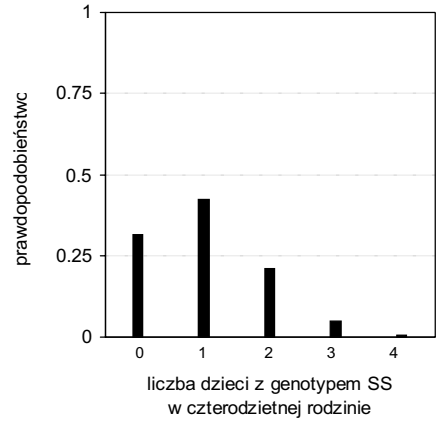
Suma prawdopodobieństw wszystkich możliwych kombinacji wynosi 1.00, to znaczy tyle, ile należałoby oczekiwać, gdybyśmy mieli całkowitą pewność, że wystąpi właśnie dana kombinacja. Rozkład dwumianowy prawdopodobieństwa (wykres gęstości rozkładu) dla $\pi = 0.25$ oraz $n = 4$ przedstawia Ryc. 16.

Do obliczeń moglibyśmy także zastosować ogólne równanie służące do liczenia prawdopodobieństw rozkładu dwumianowego:

$$P(r\ A) = \frac{n!}{r!(n-r)!}\pi^r(1-\pi)^{n-r}$$

Wykonajmy obliczenia tą metodą dla wariantu wystąpienia 2 dzieci z genotypem *SS*:

$$P(2\_SS) = \frac{4!}{2!(4-2)!} \times 0.25^2(1-0.25)^{4-2} = \frac{4\times3\times2\times1}{1\times2\times1\times2} \times 0.0625 \times 0.5625 = 0.2109375$$



Ryc. 16. Rozkład prawdopodobieństw występowania genotypu *SS* w czterodzietnej rodzinie u dzieci, których oboje rodzice są heterozygotami *Ss*. Prawdopodobieństwo wystąpienia genotypu *SS* u dziecka takich rodziców wynosi $\pi = 0.25$.

Rozkład Poissona

Przykład 20

Zobowiązania likwidowanej małej przychodni zdrowia zamierza przejąć inna większa przychodnia, która przyjmuje średnio 4.2 pacjenta na dzień, ale nie może przyjąć więcej pacjentów dziennie niż 10. Oszacowano, że po zamknięciu likwidowanej małej przychodni średnia liczba pacjentów w większej przychodni wzrośnie do 6.1 dziennie. Przez ile dni w roku istnieje niebezpieczeństwo, że liczba wizyt na dzień w większej przychodni zdrowia przekroczy 10?

Naszym zadaniem jest policzyć prawdopodobieństwo, że liczba wizyt na dzień osiągnie 11 lub więcej. W tabeli poniżej przedstawiono oszacowane prawdopodobieństwa 0, 1, 2, 3... lub 10 wizyt dziennie:

liczba wizyt	prawdopodobieństwo
0	0.0022
1	0.0137
2	0.0417
3	0.0848
4	0.1294
5	0.1579
6	0.1605
7	0.1399
8	0.1066
9	0.0723
10	0.0440
suma (0-10)	0.9530
≥ 11	1 – 0.9531 = 0.0469

liczone z równania:

$$P(x) = \frac{e^{-6.1} 6.1^{x}}{x!} \quad \text{gdzie } \mu = 6.1$$

Z różnicy mamy, że prawdopodobieństwo 11 lub więcej wizyt na dzień wynosi 0.0469 czyli 4.69%, co daje 0.0469 x 365 dni w roku = 17.12. A zatem, po przejęciu pacjentów małej przychodni przez większą przychodnię przez 17-18 dni w roku może wystąpić większa liczba wizyt niż maksymalnie dopuszczalna w większej przychodni, czyli niż 10.

Przykład 21

Próbkę kwasu acetylosalicylowego, wyznakowaną izotopowo przy węglu $^{[14]}CH_3$-COO-$C_7H_5O_2$, zliczano przez 5 minut w liczniku scyntylacyjnym. Liczba zliczeń wynosiła 2905. Jaki jest błąd standardowy oznaczenia?

Błąd wyznaczamy z równania:

$$SE = \sqrt{x} = \sqrt{2905} = 53.89$$

Jeżeli chcielibyśmy wyrazić wynik jako liczbę zliczeń na minutę (cpm) to mamy:

$$\text{cpm} = 2905/5 = 581$$

Wartość ta odpowiada

$$\lambda = \frac{\mu}{t} = \frac{2905}{5} = 581$$

oraz SE = $\sqrt{x} = 53.89/5 = 10.78 = \sqrt{\frac{\lambda}{t}} = \sqrt{\frac{581}{5}} = \sqrt{116.2} = 10.78$

95% przedział ufności dla średniej liczby zliczeń na minutę (cpm) obliczamy korzystając z aproksymacji do rozkładu normalnego:

$$581 \pm 1.96 \times 10.78 = 559.87 \text{ do } 602.13 \text{ cpm}$$

Przykład 22

Tą samą próbkę zliczono trzykrotnie i uzyskano następujące wyniki:

	liczba zliczeń	**czas (min)**
1	1799	3
2	2453	4
3	4387	8
suma	8639	15

$$\text{cpm} = 8639/15 = 575.93 \qquad SE = \frac{\sqrt{8639}}{15} = \frac{92.95}{15} = 6.20$$

Widzimy, że średnia liczba zliczeń na minutę jest podobna do tej obliczonej na podstawie wyniku zliczanego przez 5 minut, ale wartość błędu jest prawie dwukrotnie niższa, gdyż czas obserwacji był 3-krotnie dłuższy.

Rozdział 15

Zastosowania testów istotności dla pojedynczej próby

Sparowany test t

Przykład 23

U pacjentów poddawanych zabiegom z zastosowaniem krążenia pozaustrojowego badano reaktywność płytek krwi przed zabiegiem oraz w 5 dobie po zabiegu, określając ekspresję selektyny P na powierzchni płytek krwi poddanych działaniu trombiny w warunkach *in vitro*. Czy trombina wpływa na podwyższenie ekspresji selektyny P na powierzchni płytek oraz czy zabieg kardiochirurgiczny wpływa na zmniejszenie reaktywności płytek w obecności trombiny?

pacjent	płytki spocz. przed zabiegiem	płytki aktyw. trombiną przed zabiegiem	płytki spocz. po zabiegu	płytki aktyw. trombiną po zabiegu	różnica przed zabiegiem	różnica po zabiegu	różnica płytki spocz.	różnica płytki aktyw.
1	1.80	44.20	1.90	38.80	42.40	36.90	0.20	–5.40
2	2.80	73.80	3.50	73.60	71.00	70.10	0.70	–0.20
3	2.60	70.50	2.50	55.50	67.90	53.00	–0.10	–15.00
4	2.20	60.80	2.20	51.50	58.60	49.30	0.10	–9.30
5	2.00	58.70	2.70	64.10	56.70	61.40	0.60	5.40
6	1.70	52.10	1.20	30.30	50.40	29.10	–0.50	–21.80
7	1.50	45.70	1.90	49.50	44.20	47.60	0.40	3.80
8	2.10	67.20	2.10	55.70	65.10	53.60	0.00	–11.50
9	1.90	61.30	2.00	54.70	59.40	52.70	0.10	–6.60
10	2.10	71.40	1.80	52.50	69.30	50.70	–0.30	–18.90
11	1.70	61.20	1.70	51.60	59.50	49.90	0.00	–9.60
12	1.80	64.00	1.80	55.00	62.20	53.20	0.00	–9.00
13	1.60	57.40	2.10	68.50	55.80	66.40	0.60	11.10
14	1.50	55.70	1.40	45.50	54.20	44.10	–0.10	–10.20
15	1.40	54.70	1.70	58.20	53.30	56.50	0.30	3.50
16	1.30	52.50	1.80	61.80	51.20	60.00	0.50	9.30
17	1.60	63.70	0.80	29.70	62.10	28.90	–0.70	–34.00
18	1.60	67.80	1.10	42.20	66.20	41.10	–0.50	–25.60
19	1.20	53.60	1.40	54.20	52.40	52.80	0.20	0.60
20	1.10	48.50	1.60	62.30	47.40	60.70	0.50	13.80

średnia	1.77	59.24	1.87	52.76	57.47	50.89	0.09	–6.48
SD	0.44	8.48	0.59	11.35	8.17	10.93	0.40	12.68
SE	0.10	1.90	0.13	2.54	1.83	2.44	0.09	2.83
				t_{par}	31.45	20.83	1.04	–2.29
					$t_{(1,\alpha)}$	$t_{(1,\alpha)}$	$t_{(2,\alpha)}$	$t_{(1,\alpha)}$
			5% ($\alpha = 0.05$)		1.73	1.73	2.09	1.73
			1% ($\alpha = 0.01$)		2.54	2.54	2.86	2.54
			2.5% ($\alpha = 0.025$)					2.09

Nasze zadanie jest dwojakie: po pierwsze, musimy sprawdzić, czy zadziałanie trombiny na płytki wywołuje wzrost ekspresji selektyny P, a jeżeli tak, to po drugie – czy trombina „prowokuje" słabszy wzrost ekspresji selektyny P w płytkach pacjentów po przeprowadzeniu u tych pacjentów zabiegu kardiochirurgicznego.

W przypadku pierwszego pytania nasza hipoteza zerowa brzmi:

– H_0: średnia ekspresja selektyny P w płytkach aktywowanych trombiną nie jest większa niż ekspresja selektyny P w płytkach spoczynkowych,

natomiast hipoteza alternatywna brzmi:

– H_A: średnia ekspresja selektyny P w płytkach aktywowanych trombiną jest większa niż ekspresja selektyny P w płytkach spoczynkowych, czyli

$H_0: \mu_{akt} \leq \mu_{spocz}$
$H_A: \mu_{akt} > \mu_{spocz}$.

W przypadku drugiego pytania nasza hipoteza zerowa mówi, że:

– H_0: średnia ekspresja selektyny P w płytkach aktywowanych trombiną u pacjentów po zabiegu nie jest mniejsza niż średnia ekspresja selektyny P w płytkach aktywowanych trombiną u tych samych pacjentów przed zabiegiem,

hipoteza alternatywna mówi natomiast, że:

– H_A: średnia ekspresja selektyny P w płytkach aktywowanych trombiną u pacjentów po zabiegu jest mniejsza niż średnia ekspresja selektyny P w płytkach aktywowanych trombiną u tych samych pacjentów przed zabiegiem, czyli

$H_0: \mu_{po} \geq \mu_{przed}$
$H_A: \mu_{po} < \mu_{przed}$.

Zauważmy, że mamy tutaj do czynienia z ***testem jednostronnym***, gdyż weryfikujemy nie wystąpienie zmian (*in plus* bądź *in minus*), ale konkretnie wzrostu ekspresji pod wpływem trombiny lub zmniejszenia reaktywności po zabiegu. Są to ewidentnie sparowane pomiary, ponieważ: 1) trombiną działamy na te same płytki, w których sprawdzaliśmy uprzednio ekspresję w stanie spoczynkowym (parę tworzą tutaj płytki spoczynkowe czyli niepobudzone oraz płytki aktywowane), 2) reaktywność (odpowiedź płytek na działanie trombiny) przed zabiegiem i po zabiegu badamy u tych samych pacjentów.

Obliczone wartości statystyki testu *t* wynoszą:

a) dla pary płytki spoczynkowe-płytki pobudzone trombiną u pacjentów przed zabiegiem: 31.45,
b) dla pary płytki spoczynkowe-płytki pobudzone trombiną u pacjentów po zabiegu: 20.83,

c) dla pary płytki pobudzone trombiną przed zabiegiem-płytki pobudzone trombiną po zabiegu: –2.29.

Odczytane z tablic wartości statystyki testu *t* Studenta dla 20 – 1 = 19 stopni swobody wynoszą: 1.73 dla poziomu istotności 0.05 oraz 2.54 dla poziomu istotności 0.01. Oznacza to, że dla rozkładu z 19 stopniami swobody istnieje 5% prawdopodobieństwo znalezienia wartości krytycznej większej niż 1.73 oraz że istnieje 1% prawdopodobieństwo znalezienia wartości krytycznej większej niż 2.54. Innymi słowy, istnieje prawdopodobieństwo 95% lub większe, że średnia ekspresja selektyny P w płytkach aktywowanych trombiną u pacjentów po zabiegu jest mniejsza niż średnia ekspresja selektyny P w płytkach aktywowanych trombiną u tych samych pacjentów przed zabiegiem oraz istnieje prawdopodobieństwo 99% lub większe, że ekspresja selektyny P w płytkach aktywowanych trombiną jest większa niż w płytkach spoczynkowych.

Dla porównania, płytki spoczynkowe-płytki pobudzone trombiną u pacjentów przed zabiegiem wartość doświadczalna (obliczona) $t_{(1,\alpha)}$ = 31.45 jest o wiele większa od wartości tablicowych, zarówno przy poziomie istotności 0.05, jak i 0.01. Oznacza to, że jeżeli odrzucimy hipotezę zerową, to ryzyko popełnienia błędu, czyli odrzucenia prawdziwej hipotezy zerowej, wynosi mniej niż 1% (błądzimy w 1 na 100 przypadków). Innymi słowy, prawdopodobieństwo, że tak duża różnica, jak ta wykazana w toku obliczeń, mogłaby wynikać z czystego przypadku jest niższe niż 1%. Podobnie, ponieważ wartość doświadczalna $t_{(1,\alpha)}$ = 20.83 jest o wiele większa od wartości tablicowych, zarówno przy poziomie istotności 0.05, jak i 0.01, możemy stwierdzić, że różnica ekspresji selektyny P w płytkach spoczynkowych i aktywowanych u pacjentów po zabiegu jest rzeczywista i mogłaby wynikać z czystego przypadku rzadziej niż w 1 na 100 przypadków.

Sprawdzając, czy reaktywność płytek po zabiegu jest istotnie niższa, stwierdziliśmy, że wartość $t_{dośw}$ = 2.29 jest wyższa od $t_{(1)\ 0.05}$ = 1.73, czy $t_{(1)\ 0.025}$ = 2.09, ale mniejsza od $t_{(1)\ 0.01}$ = 2.54. Oznacza to, że prawdopodobieństwo, iż tak duża różnica w reaktywności przed i po zabiegu, jak ta zaobserwowana w przebadanej grupie pacjentów, wynosi mniej niż 5%, a nawet nieco mniej niż 2.5%, lecz więcej niż 1%.

Możemy zapytać: czy obserwowane zmiany w reaktywności płytek po zabiegu nie mogą wynikać z faktu, że ekspresja selektyny P w płytkach spoczynkowych jest różna u tych samych pacjentów przed i po zabiegu? *De facto*, mamy prawo oczekiwać, że zabieg będzie wpływał na stopień aktywacji płytek krążących (spoczynkowych). Gdyby tak naprawdę było, to niższy stopień aktywacji płytek krążących u pacjentów po zabiegu (niższe wartości początkowe) mógłby teoretycznie tłumaczyć ich obniżoną reaktywność (niższe wartości końcowe), natomiast wyższy stopień aktywacji płytek krążących (spoczynkowych) u pacjentów po zabiegu (wyższe wartości początkowe) pogłębiałby wydźwięk zjawiska obniżonej reaktywności (niższe wartości końcowe) po zabiegu. Powinniśmy zatem dodatkowo porównać – pomimo, że problem ten nie został sprecyzowany w zadaniu – stopień ekspresji selektyny P w płytkach spoczynkowych. Nie precyzujemy jakiego kierunku zmian oczekujemy – sprawdzamy po prostu, czy aktywacja płytek spoczynkowych przed i po zabiegu jest taka sama czy różna – dlatego możemy zastosować ***test obustronny***. Jeżeli nie wykażemy różnic, to będziemy wiedzieli, że obserwowane zmiany w reaktywności należy przypisać różnej reakcji płytek na trombinę przed i po zabiegu, jeżeli wykazalibyśmy zmiany, to wnioskowanie byłoby bardziej złożone.

Hipoteza zerowa brzmi:

– H_0: średnia ekspresja selektyny P w płytkach spoczynkowych jest równa u tych samych pacjentów przed i po zabiegu,

natomiast hipoteza alternatywna brzmi:

– H_A: średnia ekspresja selektyny P w płytkach spoczynkowych jest różna u tych samych pacjentów przed i po zabiegu, czyli:

$H_0: \mu_{przed} = \mu_{po}$
$H_A: \mu_{przed} \neq \mu_{po}$

Dla tego porównania obliczona wartość t wynosi 1.04 i jest wyraźnie mniejsza od wartości tabelarycznych dla prawdopodobieństwa 5% lub 1%. Nie mamy zatem podstaw do odrzucenia hipotezy zerowej w tym przypadku i wnioskujemy, że ekspresja selektyny P w płytkach spoczynkowych nie jest istotnie różna u tych samych pacjentów przed i po zabiegu. W teoretycznych rozważaniach na temat testu t podaliśmy, że dla danej wartości krytycznej testu t przy określonej liczbie stopni swobody odpowiadający poziom istotności dla testu jednostronnego będzie zawsze mniejszy od poziomu istotności dla testu obustronnego. Kuszące jest zadanie pytania: czy gdybyśmy sprecyzowali kierunek zmian, tzn. poddali testowaniu hipotezę mówiącą na przykład, że ekspresja selektyny P jest wyższa u tych samych pacjentów po zabiegu, uzyskalibyśmy istotną różnicę? Tablicowa wartość krytyczna testu t dla największego dopuszczalnego prawdopodobieństwa 5% (tzn. poziomu istotności 0.05) przy 19 stopniach swobody wynosi 1.33, jest zatem i tak większa od t doświadczalnego (1.04). Możemy zatem uznać, że nie ma podstaw do odrzucenia hipotezy zerowej nawet przy przyjęciu stosunkowo dużego ryzyka błędu. Średnia ekspresja selektyny P w płytkach spoczynkowych (nie aktywowanych trombiną) jest taka sama, niezależnie od tego, czy obserwacji dokonujemy przed czy po zabiegu.

Przykład 24

Szczury poddawano intensywnemu wysiłkowi fizycznemu oraz badano jaki jest wpływ wysiłku na zmianę masy ciała. Uzyskano następujące różnice mas ciała (g) przed i po wysiłku (μ) u 12 przebadanych zwierząt:

1.7	–1.4
–0.4	–1.8
0.7	–1.2
0.2	–0.9
–1.8	0.9
–1.8	–2.0

Czy wysiłek fizyczny wpływa na zmianę masy ciała u badanych zwierząt?

Hipotezy:

– H_0: $\mu = 0$ (wysiłek fizyczny nie wpływa na zmianę masy ciała, czyli średnia różnica masy ciała przed i po wysiłku wynosi 0),

– H_A: $\mu \neq 0$ (wysiłek fizyczny wpływa na zmianę masy ciała, czyli średnia różnica masy ciała przed i po wysiłku jest różna 0),

testujemy przy poziomie istotności $\alpha = 0.05$ oraz $n - 1 = 12 - 1 = 11$ stopniach swobody, czyli zakładamy, że popełnimy najwyżej 5% błędu przy odrzuceniu prawdziwej hipotezy zerowej mówiącej, że wysiłek fizyczny nie wpływa na masę ciała.

Średnia próby wynosi

$$\bar{x} = \sum x_i \Big/ n = -7.8/12 = -0.65 \text{ g},$$

zaś odchylenie $s = 1.2523$ g

$$t_{dośw} = \frac{\bar{x}}{s/\sqrt{n}} = \frac{-0.65}{1.2523/\sqrt{12}} = \frac{-0.65}{0.36} = -1.81,$$

tablicowe $t_{0.05(2),\,11} = 2.201$.

Ponieważ $|t_{\text{dośw}}| < t_{0.05(2),\,11}$, nie mamy podstaw do odrzucenia hipotezy zerowej ($0.05 < p < 0.10$).

Na podstawie zebranych obserwacji nie możemy stwierdzić, że wysiłek fizyczny wpływa na zmianę masy ciała u szczurów.

Test t Studenta dla pojedynczej próby

Przykład 25

Chcemy określić stechiometrię wiązania jonów wapnia przez białko o strukturze modelowej zsyntetyzowane metodami inżynierii genetycznej. Wiemy, że białko natywne wiąże 3 jony wapnia na cząsteczkę. Oczekujemy, że wprowadzenie nowej domeny do cząsteczki modelowej wpłynie na podwyższenie tej stechiometrii do 4 jonów na cząsteczkę. W serii eksperymentów wyznaczono (metodą Scatcharda) liczbę jonów Ca^{2+} wiązanych przez strukturę modelową i uzyskano następujące wyniki:

pomiar nr	Ca^{2+}/cząsteczkę	$\bar{x} - \mu_{(3)}$	$\bar{x} - \mu_{(4)}$
1	3.73	0.73	–0.27
2	4.23	1.23	0.23
3	4.03	1.03	0.03
4	3.69	0.69	–0.31
5	5.02	2.02	1.02
6	4.14	1.14	0.14
7	3.88	0.88	–0.12
8	2.95	–0.05	–1.05
9	4.23	1.23	0.23
10	3.82	0.82	–0.18
11	4.11	1.11	0.11
12	3.76	0.76	–0.24
średnia	3.97	0.97	–0.03
SD	0.48	0.48	0.48

Z jakim prawdopodobieństwem możemy orzec, że modyfikacja struktury białka wpłynęła na zmianę stechiometrii wiązania jonów Ca^{2+} z 3 na 4?

Pierwszym pytaniem, jakie możemy sobie zadać jest: czy zmodyfikowana struktura białka wpływa na wiązanie większej liczby jonów Ca^{2+} niż 3? Chcielibyśmy, aby prawdopodobieństwo błędnego wnioskowania nie było większe niż 0.1%, czyli dopuszczamy, że pomylimy się raz na 1000 przypadków

Łatwo zauważyć, że większość wyznaczonych wartości stechiometrycznych jest wyższa od 3 (trzecia kolumna tabeli). Naszym zadaniem jest obliczyć, na ile istotna jest to

różnica i czy obserwowane różnice nie mieszczą się w granicach naturalnej zmienności. Hipoteza zerowa i hipoteza alternatywna mają postać:

$H_0: \bar{x} - 3 \leq 0$

$H_A: \bar{x} - 3 > 0$

Wartość t (jednostronna gdyż testujemy prawdziwość nierówności $\bar{x} - 3 \leq 0$ i $\bar{x} - 3 > 0$) wyznaczona na podstawie wyników w trzeciej kolumnie wynosi:

$$t_{\text{dośw}} = \frac{\bar{x}}{s/\sqrt{n}} = \frac{0.97}{0.48/\sqrt{12}} = \frac{0.97}{0.139} = 7.0$$

Krytyczna wartość t = 4.02 dla 12 – 1 = 11 stopni swobody i poziomu istotności 0.001 jest mniejsza od $t_{\text{dośw}}$, czyli z prawdopodobieństwem nie mniejszym niż 0.1% możemy odrzucić hipotezę zerową i stwierdzić, że zmodyfikowana struktura białka wiąże większą liczbę jonów Ca^{2+} niż 3 na cząsteczkę.

Skoro wiemy już, że zmodyfikowana cząsteczka białka wiąże więcej jonów Ca^{2+} niż cząsteczka natywna, obliczymy teraz, czy nasze teoretyczne oszacowanie, iż stechiometria wiązania wynosi 4 jest prawdziwe. Weryfikacja poprzedniej pary hipotez, że na jedną cząsteczkę białka przypadają 3 jony wapnia, daje nam następujące teoretyczne warianty stechiometrii wiązania:

a) liczba jonów jest mniejsza od 4, ale większa od 3, czyli rzeczywista liczba jonów przypadająca na pojedynczą cząsteczkę białka kształtuje się między 3 a 4,
b) liczba jonów jest większa od 4 lub
c) liczba ta wynosi dokładnie 4.

Para hipotez zbudowanych do zweryfikowania tego poglądu ma postać:

– $H_0: \bar{x} - 4 = 0$ (liczba jonów na cząsteczkę białka wynosi dokładnie 4),
– $H_A: \bar{x} - 4 \neq 0$ (liczba jonów na cząsteczkę białka jest różna, to znaczy albo mniejsza niż 4, albo większa niż 4).

Obliczona statystyka t wynosi:

$$t = \frac{-0.03}{0.48/\sqrt{12}} = \frac{-0.03}{0.139} = -0.217 \ .$$

Dla 11 stopni swobody taka wartość krytyczna t (obustronna, gdyż oceniamy prawdopodobieństwo tego, że wyrażenie $\bar{x} - 4$ będzie różne od zera) odpowiada prawdopodobieństwu 16.78% (0.1678), czyli nie mamy podstaw, aby odrzucić hipotezę zerową. Zauważmy, że stwierdzenie, iż nie mamy podstaw, aby odrzucić hipotezę zerową mówiącą, że liczba jonów na cząsteczkę białka wynosi dokładnie 4, nie jest równoznaczne ze stwierdzeniem, że możemy taką hipotezę przyjąć, czyli orzec, że to co mówi hipoteza zerowa jest prawdą. Powiemy jedynie, że prawdopodobieństwo, iż postąpiliśmy słusznie nie odrzucając hipotezy mówiącej, że zmodyfikowana struktura białka wiąże dokładnie 4 jony Ca^{2+} na jedną cząsteczkę białka, wynosi 100 – 16.78% = 83.2%.

Rozdział 16

Zastosowania testów istotności dla porównywania dwóch prób

Test normalny i test t Studenta

Przykład 26

U 150 dzieci wykonano badanie krwi na zakażenie *Plasmodium falciparum*. Występowanie pasożytów stwierdzono we krwi 70 dzieci. Stężenie hemoglobiny u tych dzieci wynosiło 10.6 ± 1.4 g/100 ml krwi pełnej. U pozostałych 80 dzieci średnie stężenie hemoglobiny było wyższe i wynosiło 11.5 ± 1.3 g/100 ml. Czy wyniki te mogą wskazywać na obniżenie stężenia hemoglobiny pod wpływem zakażenia *P. falciparum*?

Obie grupy: dzieci zakażone i dzieci zdrowe są bardzo liczne, możemy zatem wykorzystać test normalny do zbadania istotności różnic. Istotność zweryfikujemy na poziomie 0.001 czyli zakładając 0.1% ryzyko popełnienia błędu przy odrzuceniu prawdziwej hipotezy zerowej, która mówi, że:

H_0: $\mu_{zakażone} \geq \mu_{zdrowe}$

Zgodnie z hipotezą alternatywną oczekujemy natomiast, że:

H_A: $\mu_{zakażone} < \mu_{zdrowe}$

$$z = \frac{10.6 - 11.5}{\sqrt{\left(1.4^2/70 + 1.3^2/80\right)}} = \frac{-0.9}{0.222} = -4.05$$

jest większa od tablicowej wartości $z_{0.001}$ = 3.29, czyli możemy wnioskować, że stężenie hemoglobiny u dzieci zakażonych *P. falciparum* jest niższe niż u dzieci zdrowych.

Przedział ufności dla tego przypadku wynosi: – 0.9 ± (1.96 x 0.222) = od –1.34 do –0.46 g/100 ml, czyli średnie stężenie hemoglobiny u dzieci zakażonych jest od 0.46 do 1.34 g/100 ml niższe niż u dzieci zdrowych.

Przykład 27

W tabeli poniżej zestawiono stężenia fibrynogenu w osoczu krwi u pacjentów z chorobą wieńcową oraz u ludzi zdrowych w tym samym przedziale wieku. Czy możemy powiedzieć, że chorobie wieńcowej towarzyszy podwyższone stężenie fibrynogenu w osoczu krwi?

	stężenie fibrynogenu (g/L)	
	pacjenci z chorobą wieńcową	osoby zdrowe
	2.71	2.76
	3.51	3.60
	3.54	3.75
	3.26	3.59
	4.13	3.63
	3.31	2.38
	3.83	2.34
	3.61	3.23
	4.08	3.52
	3.60	3.85
	3.21	3.27
	3.73	2.90
	3.60	2.84
	3.79	3.18
	3.99	
średnia	3.5933	3.2029
SD	0.3707	0.4927
n	15	14

Testujemy parę hipotez:
H_0: $\mu_{chorzy} \leq \mu_{zdrowi}$
H_A: $\mu_{chorzy} > \mu_{zdrowi}$

$$s = \sqrt{\left[\frac{\left(14 \times 0.3707^2 + 13 \times 0.4927^2\right)}{\left(15+14-2\right)}\right]} = 0.4337 \text{ g/L}$$

$$t = \frac{\left(3.5933 - 3.2029\right)}{0.4337\sqrt{\left(1/15+1/14\right)}} = \frac{0.3904}{0.1612} = 2.422,$$

dla *d.f.* = 15 + 14 – 2 = 27 stopni swobody $t_{0.01(1),\,27} > t_{dośw} > t_{0.025(1),\,27}$ czyli prawdopodobieństwo, że stężenie fibrynogenu w osoczu u pacjentów z choroba wieńcową jest wyższe niż u osób zdrowych wynosi $0.01 < p < 0.025$ ($1\% < P < 2.5\%$).

Przedział ufności dla różnicy wynosi:

$$CI(95\%) = (\bar{x}_1 - \bar{x}_2) \pm (t' \times SE), \qquad SE = s\sqrt{(1/n_1 + 1/n_2)}$$

0.3904 ± (2.05 x 0.1612) = od 0.06 do 0.72 g/L, czyli u pacjentów z choroba wieńcową stężenie fibrynogenu w osoczu jest od 0.06 do 0.72 g/L wyższe niż u zdrowych ludzi.

Testy istotności dla proporcji

Przykład 28

W badaniu klinicznym porównywano dwa preparaty uśmierzające ból: lek A i lek B. Do badania włączono 12 osób uskarżających się na bóle migrenowe i stosując metodę

pojedynczej ślepej próby badano u nich wpływ każdego leku na łagodzenie bólu, 9 pacjentów opowiedziało się za większą skutecznością leku A, podczas gdy 3 wybrało lek B. Czy możemy wnioskować, że lek A jest średnio bardziej skuteczny w łagodzeniu bóli migrenowych, czy też zaobserwowana różnica była spowodowana przez czysty przypadek?

Hipoteza zerowa zakłada, że oba leki są jednakowo skuteczne w łagodzeniu bólu. Jeżeli byłaby to prawda, to pacjent wytypowałby z jednakowym prawdopodobieństwem lek A, jak i lek B. Czyli moglibyśmy oczekiwać, że w przybliżeniu połowa pacjentów wybierze lek A, zaś druga połowa – lek B. Teoretycznym rozkładem opisującym taką sytuację byłby rozkład dwumianowy z wartościami $p = 0.5$ oraz $n = 12$. Ponieważ rozkład jest symetryczny dla $p = 0.5$, należałoby oczekiwać, że prawdopodobieństwo trzech odpowiedzi „A" jest takie samo, jak prawdopodobieństwo 9 odpowiedzi „A". To sytuacja zupełnie analogiczna do tej, z którą mielibyśmy do czynienie rzucając monetą 12 razy: oczekiwalibyśmy wtedy, że prawdopodobieństwo, iż reszka pojawi się 3 razy będzie takie samo jak prawdopodobieństwo, iż pojawi się 9 razy. Pamiętamy, że istotność statystyczna jest to prawdopodobieństwo, że wynik przyjmie wartość obserwowaną lub dowolną wartość bardziej skrajną od oczekiwanej. W naszym przypadku zatem, istotność będzie równa prawdopodobieństwu 9, 10, 11 lub 12 odpowiedzi A lub (plus) prawdopodobieństwu 3, 2, 1 lub 0 wyborów A (pamiętamy o regule addytywności w rachunku prawdopodobieństwa). Oto prawdopodobieństwa dla 9, 10, 11 lub 12 odpowiedzi A (prawa strona rozkładu) oraz dla 3, 2, 1 lub 0 odpowiedzi A (lewa strona rozkładu) obliczone wg równania:

$$P(r\ \ A) = \frac{n!}{r!(n-r)!}\pi^r(1-\pi)^{n-r}$$

odpowiedzi pacjentów preferujących lek A		**prawdopodobieństwo**
wartość obserwowana:	9	0.053711
wartości bardziej skrajne niż obserwowana:	10	0.016113
	11	0.002930
	12	0.000244
suma: 9, 10, 11, 12		0.072998
lewa strona rozkładu:	3	0.053711
	2	0.016113
	1	0.002930
	0	0.000244
suma: 3, 2, 1, 0		0.072998
obustronny poziom istotności:		**0.145996**

Zauważmy, że sumy prawdopodobieństw dla prawej i lewej strony rozkładu są równe, tak jak należałoby tego oczekiwać dla rozkładu symetrycznego z wartością $p = 0.5$. Suma prawdopodobieństw dla obu stron rozkładu wynosi 0.146, czyli wnioskujemy, iż prawdopodobieństwo uzyskania wyniku tak skrajnego lub bardziej skrajnego jak 9 odpowiedzi A na 12 ankietowanych przez zwykły przypadek wynosi blisko 15%. Ponieważ prawdopodobieństwo to jest znacznie wyższe niż 5%, należy stwierdzić, że na podstawie wyników tego badania nie jesteśmy w stanie wykazać różnic skuteczności obu leków w łagodzeniu bóli migrenowych u badanych pacjentów.

Przykład 29

Dla danych z poprzedniego przykładu wykorzystaj test istotności dla proporcji w aproksymacji normalnej rozkładu dwumianowego.

Proporcja pacjentów, którzy woleli lek A wynosiła $p = 9/12 = 0.75$ w porównaniu do teoretycznej wartości $p = 0.5$ (zgodnej z hipotezą zerową). Wartość statystyki testu normalnego dla proporcji wynosi:

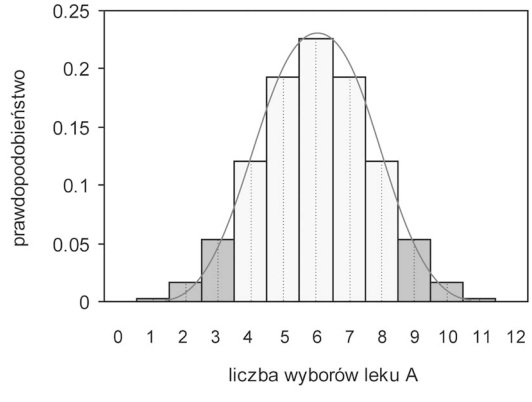
$$z = \frac{p - \pi}{\sqrt{\{\pi(1-\pi)/n\}}} = \frac{0.75 - 0.5}{\sqrt{0.5(1-0.5)/12}} = \frac{0.25}{0.1443375} = 1.732$$

Wartość prawdopodobieństwa odpowiadającego tej wartości krytycznej wynosi 0.0836 czyli 8.36%. Jak widać, jest ono prawie dwukrotnie niższe niż to oszacowane z wykorzystaniem równania na obliczanie prawdopodobieństw rozkładu dwumianowego (14.6%). Jest tak dlatego, że wykorzystaliśmy ciągły rozkład normalny do aproksymacji rozkładu dwumianowego, który jest rozkładem dyskretnym. W celu matematycznie i rachunkowo poprawniejszej i dokładniejszej aproksymacji powinniśmy zastosować poprawkę na ciągłość w równaniu testu normalnego dla proporcji:

$$z = \frac{|p - \pi| - 1/(2n)}{\sqrt{\{\pi(1-\pi)/n\}}} = \frac{|0.75 - 0.5| - 1/(24)}{\sqrt{0.5(1-0.5)/12}} = \frac{0.20833}{0.1443375} = 1.443$$

Dla takiej wartości krytycznej prawdopodobieństwo wynosi 2 x 0.0749 = 0.1498 czyli 14.98% – jest to o wiele bliższa aproksymacja do wartości 14.6% obliczonej dla rozkładu dwumianowego.

Statystyczny sens stosowania poprawki na ciągłość pokazuje Ryc. 17. Stosując poprawkę obliczamy prawdopodobieństwa dla $r = 3.5$ oraz $r = 8.5$, zamiast dla $r = 3$ i $r = 9$, co daje wyższą zgodność obu rozkładów.



Ryc. 17. Rozkład dwumianowy przy $n = 12$ i $\pi = 0.5$ dla różnej liczby wyborów leku A aproksymowany do rozkładu normalnego. Ciemniejsze pola oznaczają prawdopodobieństwa dla liczby wyborów tak skrajnych lub bardziej skrajnych niż 9.

Przykład 30

Dla danych z poprzedniego przykładu należy policzyć wartość błędu standardowego oraz przedział ufności w aproksymacji do rozkładu normalnego.

Błąd standardowy wynosi:

$$SE = \sqrt{\{p(1-p)/n\}} = \sqrt{0.75(1-0.75)/12} = 0.125$$

Przedział ufności dla 95% ($z' = 1.96$):

$$CI = p \pm z' \times SE = 0.75 \pm 1.96 \times 0.125 = 0.75 \pm 0.245 = 0.505, 0.995$$

Prawdopodobieństwo wyboru leku A przez pacjentów wynosi od 50.5% do 99.5%.

Przykład 31

23 z 251 osób ($p_1 = r_1 / n_1 =$ 0.0916 czyli 9.16%), które zostały zaszczepione przeciw grypie, zachorowało na grypę pomimo szczepienia, natomiast wśród 214 osób nie szczepionych zachorowało 91 ($p_2 = r_2 / n_2 =$ 0.425 czyli 42.5%). Czy obserwacja taka stanowi przekonujący dowód na to, że szczepionka przeciw grypie jest skuteczna?

Całkowita proporcja osób, które zachorowały na grypę, niezależnie od tego czy były szczepione czy nie, wynosi:

$$p = \frac{r_1 + r_2}{n_1 + n_2} = \frac{23 + 91}{251 + 214} = \frac{114}{469} = 0.243 \quad \text{czyli} \quad 24.3\%$$

W ten sposób:

$$z = \frac{|p_1 - p_2| - \{1/(2n_1) + 1/(2n_2)\}}{\sqrt{\{p(1-p)(1/n_1 + 1/n_2)\}}} = \frac{|0.0916 - 0.425| - \{1/(2\times 251) + 1/(2\times 214)\}}{\sqrt{\{0.243\times(1-0.243)(1/251 + 1/214)\}}} =$$

$$= \frac{0.3334 - (0.001992 + 0.0023364)}{\sqrt{0.183951 \times 0.008657}} = \frac{0.3291}{0.0399056} = 8.247$$

Ponieważ wartość $z = 8.247 > z_{0.0001} = 3.99$, możemy z prawdopodobieństwem ponad 99.99% wnioskować, że szczepionka jest skuteczna, gdyż osoby zaszczepione chorują istotnie rzadziej.
Różnica częstości zachorowań wynosi:

$$p_1 - p_2 = 0.425 - 0.0916 = 0.3334$$

Obliczmy jeszcze przedział ufności dla tej różnicy:

$$SE = \sqrt{\{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2\}} =$$

$$= \sqrt{\{0.0916 \times (1 - 0.0916) / 251 + 0.425 \times (1 - 0.425) / 214\}} = 0.03839$$

Wartość 95% przedziału ufności (z = 1.96) wynosi:

$$95\%CI = (p_1 - p_2) \pm (z' \times SE) = -0.3334 \pm (1.96 \times 0.03839) = -0.4086, -0.2592$$

Prawdopodobieństwo zachorowania na grypę jest o 25.9% do 40.9% niższe wśród osób, które otrzymały szczepionkę.

Rozdział 17

Analiza wariancji, testy porównań wielokrotnych oraz testy do oceny zgodności i biozgodności

Przykład 32

Badano wpływ polimorfizmu genetycznego glikoproteiny GPIIIa receptora dla fibrynogenu na stopień agregacji płytek we krwi pełnej pod wpływem 2 mg/ml kolagenu. Do badań włączono 41 osób reprezentujących następujące warianty genotypowe: homozygoty $Pl^{A1/A1}$ („typ dziki"), heterozygoty $Pl^{A1/A2}$ oraz homozygoty $Pl^{A2/A2}$. Zebrane wyniki przedstawiono w tabeli:

genotyp $Pl^{A1/A2}$	**liczebność** (n_i)	**średnia** ($\bar{x}_i$)	**odch. std.** (s_i)	**agregacja płytek z kolagenem (*jedn. umowne*) wartości obserwowane (x)**			
$Pl^{A1/A1}$	20	11.0688	1.1257	9.0	9.6	10.0	10.1
				10.4	10.5	10.5	10.6
				10.8	10.9	11.4	11.4
				11.4	12.3	12.6	12.9
				10.1	11.5	12.5	13.0
$Pl^{A1/A2}$	15	14.8333	1.1531	13.3	13.6	13.9	14.9
				15.0	15.1	15.8	14.6
				16.6	15.0	17.4	13.4
				14.1	15.4	14.5	
$Pl^{A2/A2}$	6	15.6458	1.1164	14.4	14.8	15.1	15.8
				16.6	17.3		

Czy stopień agregacji płytek we krwi pełnej pod wpływem kolagenu jest jednakowy u nosicieli różnych genotypów $Pl^{A1/A2}$?

H_0: $\mu_1 = \mu_2 = \mu_3$
H_A: H_0 jest fałszywa

Obliczenia:

$N = \sum n_i = 20 + 15 + 6 = 41$, liczba grup ($k$) = 3

$\sum x = 9.0 + ... + 13.0 + 13.3 + ... + 14.5 + 14.4 + ... + 17.3 = 537.0$

$\sum x^2 = 9.0^2 + ... + 13.0^2 + 13.3^2 + ... + 14.5^2 + 14.4^2 + ... + 17.3^2 = 7268.4$

zmienność całkowita:

$SS_{\text{całk}} = \sum (x - \bar{x})^2 = \sum x^2 - \left(\sum x\right)^2 / N = 7268.4 - (537)^2/41 = 235.01,\ d.f. = N - 1 = 40$

zmienność międzygrupowa:

$SS_{\text{efektu}} = \sum n_i (\bar{x}_i - \bar{x})^2$ lub $\sum n_i (\bar{x}_i)^2 - \left(\sum x\right)^2 / N =$

$= 20 \text{ x } 11.0688^2 + 15 \text{ x } 14.8333^2 + 6 \text{ x } 15.6458^2 - (537)^2/41 = 7219.515 - 7033.39 = 176.09,$

$d.f. = k - 1 = 2$

zmienność wewnątrzgrupowa:

$SS_{\text{błędu}} = \sum (n_i - 1) s_i^2 = 19 \text{ x } 1.1257^2 + 14 \text{ x } 1.1531^2 + 5 \text{ x } 1.1164^2 = 48.924,$

$d.f. = N - k = 41 - 3 = 38$

Tabela z błędami średniokwadratowymi i statystyką testu *F*.

zmienność	SS	*d.f.*	$MS = \frac{SS}{d.f.}$	$F = \frac{MS_{\text{międzygrupowa}}}{MS_{\text{wewnątrzgrupowa}}}$
międzygrupowa	176.09	2	88.045	68.39, $p < 0.001$
wewnątrzgrupowa	48.92	38	1.287	
całkowita	235.01	40		

Ponieważ wartość krytyczna $F_{0.0005.2.38} = 9.25$ jest o wiele niższa od obliczonej $F_{\text{dośw}} = 68.4$, możemy z prawdopodobieństwem przynajmniej 99.95% odrzucić hipotezę zerową i przyjąć hipotezę alternatywną mówiącą, że stopień agregacji płytek we krwi pełnej pod wpływem kolagenu nie jest jednakowy u nosicieli różnych genotypów $Pl^{A1/A2}$.

Dwuczynnikowa analiza wariancji z równą liczbą powtórzeń w grupach

Przykład 33

W 10-osobowych grupach zdrowych ochotników oraz pacjentów z chorobą niedokrwienną serca badano jaki wpływ ma zażywanie kwasu acetylosalicylowego (aspiryny, ASA) na reaktywność płytek krwi w obecności kolagenu. Połowa osób w każdej grupie otrzymywała po 150 mg ASA dziennie w ciągu 7 dni, druga połowa otrzymywała *placebo*. Wyniki reaktywności (*wyrażone w jedn. umownych*) przedstawiono w tabeli poniżej.

Postawione poniżej hipotezy postanowiono zweryfikować przy poziomie istotności $\alpha = 0.05$:

- H_0: ASA nie wywiera wpływu na reaktywność płytek, czyli $\mu_{placebo} = \mu_{ASA}$
- H_A: ASA wpływa na reaktywność płytek, czyli $\mu_{placebo} \neq \mu_{ASA}$
- H_0: nie występują różnice w reaktywności płytek krwi między zdrowymi osobnikami a pacjentami z chorobą niedokrwienną serca, czyli $\mu_{zdrowi} = \mu_{ChNS}$
- H_A: występują różnice w reaktywności płytek krwi między zdrowymi osobnikami a pacjentami z chorobą niedokrwienną serca, czyli $\mu_{zdrowi} \neq \mu_{ChNS}$
- H_0: nie ma interakcji między podawaniem ASA a obecnością ChNS
- H_A: występuje interakcja między podawaniem ASA a obecnością ChNS

Obliczenia:

	czynnik 2				**wpływ ASA**
czynnik 1	**ChNS**		**zdrowi**		**sumy i średnie**
bez ASA	21.3	453.7	28.8	829.4	
	40.2	1616.0	29.3	858.5	
	26.2	686.4	32.0	1024.0	
	39.1	1528.8	25.0	625.0	
	35.8	1281.6	23.8	566.4	
$\text{suma}_{\text{bez ASA}}$	$\sum_{l=1}^{n} X_{1l} = 162.6$		$\sum_{l=1}^{n} X_{12l} = 138.9$		$\sum_{j=1}^{b}\sum_{l=1}^{n} X_{1jl} = 301.5$
$\text{średnia}_{\text{bez ASA}}$	$\overline{X}_{11} = 32.52$		$\overline{X}_{12} = 27.78$		$\overline{X}_{1} = 30.15$
$SS_{\text{bez ASA}}$	$\sum_{l=1}^{n} X_{11l}^2 = 5566.62$		$\sum_{l=1}^{n} X_{12l}^2 = 3903.37$		
z ASA	14.0	196.0	11.0	121.0	
	12.8	163.8	10.0	100.0	
	18.4	338.6	14.3	204.5	
	12.7	161.3	14.5	210.3	
	16.5	272.3	10.8	116.6	
$\text{suma}_{\text{z ASA}}$	$\sum_{l=1}^{n} X_{21l} = 74.4$		$\sum_{l=1}^{n} X_{22l} = 60.6$		$\sum_{j=1}^{b}\sum_{l=1}^{n} X_{2jl} = 135.0$
$\text{średnia}_{\text{z ASA}}$	$\overline{X}_{21} = 14.88$		$\overline{X}_{22} = 12.12$		$\overline{X}_{2} = 13.50$
$SS_{\text{z ASA}}$	$\sum_{l=1}^{n} X_{21l}^2 = 1131.94$		$\sum_{l=1}^{n} X_{22l}^2 = 752.38$		
wpływ grupy: sumy i średnie					
$\text{suma}_{\text{całk.}}$	$\sum_{i=1}^{a}\sum_{l=1}^{n} X_{i1l} = 237.0$		$\sum_{i=1}^{a}\sum_{l=1}^{n} X_{i2l} = 199.50$		$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{l=1}^{n} X_{ijl} = 436.5$
$\text{średnia}_{\text{całk.}}$	$\overline{X}_{1} = 23.70$		$\overline{X}_{2} = 19.95$		$\overline{X} = 21.825$

$$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{l=1}^{n} X_{ijl} = 436.5$$

$$\sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{l=1}^{5} X_{ijl}^2 = 11354.31$$

a = liczba grup działania leku = 2
b = liczba grup osobników poddanych działaniu leku = 2
n = liczba powtórzeń w grupie = 5
$N = abn = (2)(2)(5) = 20$
suma całkowita dla grupy bez ASA:

$$\sum_{j=1}^{2}\sum_{l=1}^{5} X_{1jl} = 162.6 + 138.9 = 301.5$$

suma całkowita dla grupy z ASA:

$$\sum_{j=1}^{2}\sum_{l=1}^{5} X_{2jl} = 74.4 + 60.6 = 135.0$$

suma całkowita dla grupy zdrowych:

$$\sum_{i=1}^{2}\sum_{l=1}^{5} X_{i2l} = 60.6 + 138.9 = 199.5$$

suma całkowita dla grupy z ChNS:

$$\sum_{i=1}^{2}\sum_{l=1}^{5} X_{i1l} = 74.4 + 162.6 = 237.0$$

$$C = \frac{\left(\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{l=1}^{n} X_{ijl}\right)^2}{N} = \frac{(436.5)^2}{20} = 9526.6125$$

$$SS_{\text{całk}} = \sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{l=1}^{5} X_{ijl}^2 - C = 11354.31 - 9526.6125 = 1827.6975; \qquad d.f._{\text{całk.}} = N - 1 = 19$$

$$SS_{\text{grup}} = \sum_{i=1}^{a}\sum_{j=1}^{b} \frac{\left(\sum_{l=1}^{n} X_{ijl}\right)^2}{n} - C = \frac{(74.4)^2 + (60.6)^2 + (162.6)^2 + (138.9)^2}{5} - 9526.6125 =$$

$$= 1461.3255; \qquad d.f._{\text{grup.}} = ab - 1 = (2)(2) - 1 = 3$$

$$SS_{\text{wewnątrzgrupowa/błędu}} = SS_{\text{całk.}} - SS_{\text{grup}} = 1827.6975 - 1461.3255 = 366.3720$$

$$d.f._{\text{wewnątrzgrupowa/błędu}} = ab(n-1) = (2)(2)(4) = 16$$

$$SS_{\text{czynnika A (ASA)}} = \frac{\sum_{i=1}^{a}\left(\sum_{j=1}^{b}\sum_{l=1}^{n} X_{ijl}\right)^2}{bn} - C = \frac{(suma_{ASA(-)})^2 + (suma_{ASA(+)})^2}{n_{ASA}} - C =$$

$$= \frac{(135.0)^2 + (301.5)^2}{(2)(5)} - 9526.6125 = 1386.1125; \qquad d.f._{\text{czynnika A}} = a - 1 = 1$$

$$SS_{\text{czynnika B (ChNS)}} = \frac{\sum_{j=1}^{b}\left(\sum_{i=1}^{a}\sum_{l=1}^{n} X_{ijl}\right)^2}{an} - C = \frac{(suma_{ChNS})^2 + (suma_{zdrowi})^2}{n_{grupa}} - C =$$

$$= \frac{(237.0)^2 + (199.5)^2}{(2)(5)} - 9526.6125 = 70.3125$$

$$d.f._{\text{czynnika B}} = b - 1 = 1$$

$$SS_{\text{interakcji A x B}} = SS_{\text{grup}} - SS_{\text{czynnik A}} - SS_{\text{czynnik B}} = 1461.3255 - 1386.1125 - 70.3125 = 4.901$$

$$d.f._{\text{interakcji A x B}} = (d.f._{\text{czynnika A}}) \times (d.f._{\text{czynnika B}}) = (1)(1) = 1$$

zmienność	***SS***	***d.f.***	***MS***
całkowita	1827.6975	19	
efekt główny		3	
ASA	1386.1125	1	1386.113
grupa	70.3125	1	70.3125
interakcje			
ASA x grupa	4.9005	1	4.9005
wewnątrzgrupowa (*błędu*)	366.372	16	22.89825

Dla H_0: ASA nie wywiera wpływu na reaktywność płytek.

$$F = \frac{MS_{ASB}}{MS_{błędu}} = \frac{1386.1125}{22.8982} = 60.5$$

$F_{0.05(1),1.16} = 4.49$, dlatego odrzucamy H_0; ponieważ obliczona wartość $F = 60.5$ jest o wiele wyższa od tablicowej, możemy to zrobić z bardzo wysokim prawdopodobieństwem:

$$p < 0.0005.$$

Wniosek

ASA bardzo istotnie wpływa na reaktywność płytek badanych w warunkach doświadczenia.

Dla H_0: nie występują różnice w reaktywności płytek krwi między zdrowymi osobnikami a pacjentami z chorobą niedokrwienną serca.

$$F = \frac{MS_{grupa}}{MS_{błędu}} = \frac{70.3125}{22.8982} = 3.1$$

$F_{0.05(1),1.16} = 4.49$, dlatego nie odrzucamy H_0; $p < 0.10$.

Wniosek

Grupa do jakiej należy badana osoba nie ma istotnego wpływu na reaktywność płytek badaną w warunkach doświadczenia.

Dla H_0: nie ma interakcji między podawaniem ASA a obecnością ChNS.

$$F = \frac{MS_{AxB}}{MS_{błędu}} = \frac{4.901}{22.8982} = 0.215$$

$F_{0.05(1),1.16} = 4.49$, dlatego nie odrzucamy H_0; $p > 0.50$.

Wniosek

Nie ma istotnej interakcji między podawaniem ASA a występowaniem ChNS w kształtowaniu odpowiedzi płytek na działanie kolagenu w warunkach doświadczenia.

Przykład 34

U pacjentów z chorobą niedokrwienną serca badano wpływ preparatów obniżających stężenie cholesterolu w osoczu: statyn oraz fibratów. Spośród osób biorących udział w badaniu wylosowano dwie 8-osobowe grupy pacjentów, którzy przeszli niepełnościenny lub pełnościenny zawał mięśnia sercowego. W celu określenia wpływu każdego z leków u każdego z pacjentów kontynuowano terapię hipolipemizującą przez miesiąc. Uzyskane wyniki wskazywały, że każdy z preparatów powodował obniżenie stężenia cholesterolu całkowitego w osoczu. W tabeli poniżej przedstawiono względne zmiany stężenia cholesterolu (%) w badanych grupach pacjentów.

Postawione poniżej hipotezy postanowiono zweryfikować przy poziomie istotności $\alpha = 0.05$:

- H_0: nie występują różnice między skutecznością hipolipemizującego działania statyn a skutecznością hipolipemizującego działania fibratów u leczonych pacjentów, czyli $\mu_{fibraty} = \mu_{statyny}$,
- H_A: występują różnice między skutecznością hipolipemizującego działania statyn a skutecznością hipolipemizującego działania fibratów u leczonych pacjentów, czyli $\mu_{fibraty} \neq \mu_{statyny}$,

- H_0: nie występują różnice w skuteczności hipolipemizującego działania statyn i fibratów między pacjentami, którzy przeszli zawał niepełnościeny a pacjentami po zawale pełnościennym, czyli $\mu_{niepełnościenny} = \mu_{pełnościenny}$,
- H_A: występują różnice w skuteczności hipolipemizującego działania statyn i fibratów między pacjentami, którzy przeszli zawał niepełnościeny a pacjentami po zawale pełnościennym, czyli $\mu_{niepełnościenny} \neq \mu_{pełnościenny}$,

- H_0: nie ma interakcji między rodzajem zawału u pacjentów a rodzajem leku hipolipemizującego,

– H_A: występuje interakcja między rodzajem zawału u pacjentów a rodzajem leku hipolipemizującego.

Obliczenia:

	czynnik 2				
czynnik 1	**zawał niepełnościenny**		**zawał pełnościenny**		**rodzaj leku sumy i średnie**
statyny	56	3136	38	1444	
	37	1369	57	3249	
	51	2601	47	2209	
	44	1936	52	2704	
$\text{suma}_{\text{statyny}}$	$\sum_{l=1}^{n} X_{11l} = 188.0$		$\sum_{l=1}^{n} X_{12l} = 194.0$		$\sum_{j=1}^{b}\sum_{l=1}^{n} X_{1jl} = 382.0$
$\text{średnia}_{\text{statyny}}$	$\overline{X}_{11} = 47.0$		$\overline{X}_{12} = 48.5$		$\overline{X}_{1} = 47.75$
$\text{SS}_{\text{statyny}}$	$\sum_{l=1}^{n} X_{11l}^2 = 9042$		$\sum_{l=1}^{n} X_{12l}^2 = 9606$		
n_{statyny}	$n_{11} = 4$		$n_{12} = 4$		
fibraty	48	2304	45	2025	
	29	841	26	676	
	40	1600	34	1156	
	37	1369	38	1444	
$\text{suma}_{\text{fibraty}}$	$\sum_{l=1}^{n} X_{21l} = 154.0$		$\sum_{l=1}^{n} X_{22l} = 143.0$		$\sum_{j=1}^{b}\sum_{l=1}^{n} X_{2jl} = 297.0$
$\text{średnia}_{\text{fibraty}}$	$\overline{X}_{21l} = 38.5$		$\overline{X}_{22} = 35.75$		$\overline{X}_{2} = 37.125$
$\text{SS}_{\text{fibraty}}$	$\sum_{l=1}^{n} X_{21l}^2 = 6114$		$\sum_{l=1}^{n} X_{22l}^2 = 5301$		
n_{fibraty}	$n_{21} = 4$		$n_{22} = 4$		
wpływ grupy: sumy i średnie					
$\text{suma}_{\text{całk.}}$	$\sum_{i=1}^{a}\sum_{l=1}^{n} X_{i1l} = 342.0$		$\sum_{i=1}^{a}\sum_{l=1}^{n} X_{i2l} = 337.0$		$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{l=1}^{n} X_{ijl} = 679.0$
$\text{średnia}_{\text{całk}}$	$\overline{X}_{1} = 42.75$		$\overline{X}_{2} = 42.13$		$\overline{X} = 42.4375$

$$\sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{l=1}^{5} X_{ijl} = 679.0$$

$$\sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{l=1}^{5} X_{ijl}^2 = 30063.0$$

a = liczba grup działania leku = 2
b = liczba grup osobników poddanych działaniu leku = 2
n = liczba powtórzeń w grupie = 4
$N = abn = (2)(2)(4) = 16$

suma całkowita dla grupy statyny:

$$\sum_{j=1}^{2}\sum_{l=1}^{5} X_{1jl} = 382.0$$

suma całkowita dla grupy fibraty:

$$\sum_{j=1}^{2}\sum_{l=1}^{5} X_{2jl} = 297.0$$

suma całkowita dla grupy zawał niepełnościenny:

$$\sum_{i=1}^{2}\sum_{l=1}^{5} X_{i2l} = 342.0$$

suma całkowita dla grupy zawał pełnościenny:

$$\sum_{i=1}^{2}\sum_{l=1}^{5} X_{i1l} = 337.0, \qquad C = \frac{\left(\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{l=1}^{n} X_{ijl}\right)^2}{N} = \frac{(679)^2}{16} = 28815.063$$

$$\mathrm{SS}_{\text{całk}} = \sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{l=1}^{5} X_{ijl}^2 - C = 30063 - 28815.063 = 1247.9375; \qquad d.f._{\text{całk.}} = N - 1 = 15$$

$$\mathrm{SS}_{\text{grup}} = \sum_{i=1}^{a}\sum_{j=1}^{b}\frac{\left(\sum_{l=1}^{n} X_{ijl}\right)^2}{n} - C = \frac{(188)^2 + (194)^2 + (154)^2 + (143)^2}{4} - 28815.063 =$$

$$= 471.1875; \quad d.f._{\text{grup.}} = ab - 1 = (2)(2) - 1 = 3$$

$$\mathrm{SS}_{\text{wewnątrzgrupowa/błędu}} = SS_{\text{całk.}} - SS_{\text{grup}} = 1247.9375 - 471.1875 = 776.75$$

$$d.f._{\text{wewnątrzgrupowa/błędu}} = ab(n-1) = (2)(2)(3) = 12$$

$$\mathrm{SS}_{\text{czynnika A (lek)}} = \frac{\sum_{i=1}^{a}\left(\sum_{j=1}^{b}\sum_{l=1}^{n} X_{ijl}\right)^2}{bn} - C = \frac{(suma_{statyny})^2 + (suma_{fibraty})^2}{n_{statyny/\,fibraty}} - C =$$

$$= \frac{(382)^2 + (297)^2}{(2)(4)} - 28815.063 = 451.5625; \quad d.f._{\text{czynnika A}} = a - 1 = 1$$

$$\mathrm{SS}_{\text{czynnika B (typ zawału)}} = \frac{\sum_{j=1}^{b}\left(\sum_{i=1}^{a}\sum_{l=1}^{n} X_{ijl}\right)^2}{an} - C = \frac{(suma_{z.niep.})^2 + (suma_{z.pełność})^2}{n_{grupa}} - C =$$

$$= \frac{(342)^2 + (337)^2}{(2)(4)} - 28815.063 = 1.5625; \quad d.f._{\text{czynnika B}} = b - 1 = 1$$

$$SS_{\text{interakcji A x B}} = SS_{\text{grup}} - SS_{\text{czynnik A}} - SS_{\text{czynnik B}} = 471.1875 - 451.5625 - 1.5625 = 18.0625$$

$$d.f._{\text{interakcji A x B}} = (d.f._{\text{czynnika A}}) \text{ x } (d.f._{\text{czynnika B}}) = (1)(1) = 1$$

zmienność	***SS***	***d.f.***	***MS***
całkowita	1247.9375	15	
efekt główny		3	
rodzaj leku	451.5625	1	451.5625
rodzaj zawału	1.5625	1	1.5625
interakcje			
lek *x* zawał	18.0625	1	18.0625
wewnątrzgrupowa (błędu)	776.7500	12	64. 7292

Dla H_0: nie występują różnice między skutecznością hipolipemizującego działania statyn a skutecznością hipolipemizującego działania fibratów u leczonych pacjentów.

$$F = \frac{MS_{lek}}{MS_{błędu}} = \frac{451.5625}{64.7292} = 6.976$$

$F_{0.05(1),1.12} = 4.75$, dlatego odrzucamy H_0; możemy to zrobić nawet z wyższym prawdopodobieństwem: $p < 0.025$.

Wniosek

Statyny i fibraty istotnie różnią sie skutecznością obniżania stężenia cholesterolu u pacjentów z chorobą niedokrwienną, którzy przeszli zawał mięśnia sercowego.

Dla H_0: nie występują różnice w skuteczności hipolipemizującego działania statyn i fibratów między pacjentami, którzy przeszli zawał niepełnościenny a pacjentami po zawale pełnościennym.

$$F = \frac{MS_{zawał}}{MS_{błędu}} = \frac{1.5625}{64.7292} = 0.024$$

$F_{0.05(1),1.12} = 4.75$, dlatego nie odrzucamy H_0.

Wniosek

Rodzaj zawału nie wpływa na skuteczność działania leków.

Dla H_0: nie ma interakcji między rodzajem zawału u pacjentów a rodzajem leku hipolipemizującego.

$$F = \frac{MS_{AxB}}{MS_{błędu}} = \frac{18.0625}{64.7292} = 0.279$$

$F_{0.05(1),1.12} = 4.75$, dlatego nie odrzucamy H_0.

Wniosek

Nie ma istotnej interakcji między rodzajem podawanego leku a rodzajem zawału u badanych pacjentów.

Dwuczynnikowa analiza wariancji z nierówną liczbą powtórzeń w grupach

Przykład 35

W trzech grupach ochotników różniących się genotypem polimorfizmu glikoproteiny IIIa płytkowego receptora dla fibrynogenu ($Pl^{A1/A1}$, $Pl^{A1/A2}$ i $Pl^{A2/A2}$) badano wpływ 4 różnych blokerów tego receptora na reaktywność płytek krwi w obecności ADP. Wyniki hamowania reaktywności płytek przez każdy z badanych blokerów (%) przedstawiono w tabeli poniżej.

Postawione poniżej hipotezy postanowiono zweryfikować przy poziomie istotności $\alpha = 0.05$:

- H_0: Blokery nie różnią się pod względem hamowania reaktywności płytek w obecności ADP, czyli $\mu_1 = \mu_2 = \mu_3 = \mu_4$,
- H_A: Blokery różnią się pod względem hamowania reaktywności płytek w obecności ADP, czyli nie jest prawdziwe równanie $\mu_1 = \mu_2 = \mu_3 = \mu_4$,

- H_0: nie występują różnice w reaktywności płytek krwi między nosicielami różnych genotypów polimorfizmu $Pl^{A1/A2}$, czyli $\mu_{PlA1/A1} = \mu_{PlA1/A2} = \mu_{PlA2/A2}$,
- H_A: występują różnice w reaktywności płytek krwi między nosicielami różnych genotypów polimorfizmu $Pl^{A1/A2}$, czyli równanie $\mu_{PlA1/A1} = \mu_{PlA1/A2} = \mu_{PlA2/A2}$ nie jest prawdziwe,

- H_0: nie ma interakcji między rodzajem blokera a rodzajem genotypu,
- H_A: występuje interakcja między rodzajem blokera a rodzajem genotypu.

Obliczenia:

	czynnik 2								
czynnik 1	**bloker 1** (x_1)	x_1^2	**bloker 2** (x_2)	x_2^2	**bloker 3** (x_3)	x_3^2	**bloker 4** (x_4)	x_4^2	**sumy i średnie wpływ Pl$^{A1/A2}$**
Pl$^{A1/A1}$	11.8	139.2	14.2	201.6	14.6	213.2	9.6	92.2	
	10.9	118.8	11.6	134.6	9.5	90.3	10.1	102.0	
	14.5	210.3	10.3	106.1	8.9	79.2	7.7	59.3	
			11.9	141.6	9.1	82.8	11.1	123.2	
			11.0	121.0	11.1	123.2	7.5	56.3	
			12.2	148.8	9.6	92.2	8.4	70.6	
					10.5	110.3			
					11.0	121.0			
					11.9	141.6			
$\sum x$	37.20		71.20		96.20		54.40	$\sum x$	259.00
$\bar{x}$	12.40		11.87		10.69		9.07	$\bar{x}$	10.79
$\sum x^2$	468.3		853.74		1053.66		503.48	$\sum x^2$	2879.18
n	3		6		9		6	n	24
Pl$^{A1/A2}$	13.4	179.6	9.8	96.0	14.6	213.2	7.4	54.8	
	11.7	136.9	11.3	127.7	8.6	74.0	7.2	51.8	
	11.1	123.2	8.4	70.6	8.9	79.2	8.4	70.6	
	10.7	114.5	11.4	130.0	9.8	96.0	11.1	123.2	
			12.8	163.8	7.8	60.8	9.2	84.6	
			10.2	104.0	8.0	64.0	6.7	44.9	
			12.1	146.4	10.6	112.4	11.5	132.3	
			10.1	102.0	8.5	72.3	8.8	77.4	
					9.0	81.0			
					11.5	132.3			
					8.1	65.6			
					9.2	84.6			
$\sum x$	46.90		86.10		114.60		70.30	$\sum x$	317.90
$\bar{x}$	11.73		10.76		9.55		8.79	$\bar{x}$	9.93
$\sum x^2$	554.15		940.55		1135.32		639.59	$\sum x^2$	3269.61
n	4		8		12		8	n	32
Pl$^{A2/A2}$	14.8	219.0	17.8	316.8	12.0	144.0	15.6	243.4	
	19.3	372.5	18.9	357.2	13.9	193.2	12.2	148.8	
			11.4	130.0	14.1	198.8	10.1	102.0	
			14.9	222.0	22.1	488.4	17.4	302.8	
					16.2	262.4		0.0	
					13.4	179.6		0.0	
$\sum x$	34.10		63.00		91.70		55.30	$\sum x$	244.10
$\bar{x}$	17.05		15.75		15.28		13.83	$\bar{x}$	15.26
$\sum x^2$	591.53		1026.02		1466.43		796.97	$\sum x^2$	3880.95
n	2		4		6		4	n	16
$\sum_{i=1}^{a}\sum_{l=1}^{n} x$	118.20		220.30		302.50		180.00	$\sum x_{calk.}$	821.00
n_{bloker}	9		18		27		18	N	72
$\overline{X}$	13.13		12.24		11.20		10.00	$\bar{x}_{calk.}$	11.40

$$\sum_{i=1}^{3}\sum_{j=1}^{4}\sum_{l=1}^{n} X_{ijl} = 821.0$$

$$\sum_{i=1}^{3}\sum_{j=1}^{4}\sum_{l=1}^{n} X_{ijl}^2 = 10029.74$$

a = liczba grup genotypowych = 3
b = liczba blokerów = 4
N = całkowita liczba obserwacji = 72
suma całkowita dla

$\mathrm{Pl}^{\mathrm{A1/A1}} = \sum_{j=1}^{4}\sum_{l=1}^{n} X_{1jl} = 259.0$

suma całkowita dla

$\mathrm{Pl}^{\mathrm{A1/A2}} = \sum_{j=1}^{4}\sum_{l=1}^{n} X_{2jl} = 317.9$

suma całkowita dla

$\mathrm{Pl}^{\mathrm{A2/A2}} = \sum_{j=1}^{4}\sum_{l=1}^{n} X_{3jl} = 244.1$

suma całkowita dla

blokera 1 = $\sum_{i=1}^{3}\sum_{l=1}^{n} X_{i1l} = 118.2$

suma całkowita dla

blokera 2 = $\sum_{i=1}^{3}\sum_{l=1}^{n} X_{i2l} = 220.3$

suma całkowita dla

blokera 3 = $\sum_{i=1}^{3}\sum_{l=1}^{n} X_{i3l} = 302.5$

suma całkowita dla

blokera 4 = $\sum_{i=1}^{3}\sum_{l=1}^{n} X_{i4l} = 180.0$

$$C = \frac{\left(\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{l=1}^{n_{ij}} X_{ijl}\right)^2}{N} = \frac{821^2}{72} = 9361.6806$$

$SS_{\text{całk}} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{l=1}^{n_{ij}} X_{ijl}^2 - C = 10029.74 - 9361.6806 = 668.0594$; $d.f._{\text{całk}} = N - 1 = 71$

$$SS_{\text{grup}} = \sum_{i=1}^{a}\sum_{j=1}^{b}\frac{\left(\sum_{l=1}^{n_{ij}} X_{ijl}\right)^2}{n_{ij}} - C = 461.28 + 844.907 + 1028.27 + 493.227) + (549.903 + 926.65 +$$

$$1094.43 + 617.76) + (581.41 + 992.25 + 1401.48 + 764.52) - 9361.68 = (2827.68 + 3188.75 +$$

$$3739.66) - 9361.68 = 9756.09 - 9361.68 = 394.41;$$

$$d.f._{\text{grup}} = ab - 1 = (3)(4) - 1 = 11$$

$$SS_{\text{czynnik A}} = \sum_{i=1}^{a}\frac{\left(\sum_{j=1}^{b}\sum_{l=1}^{n_{ij}} X_{ijl}\right)^2}{\sum_{j=1}^{b} n_{ij}} - C = (1552.36 + 2696.23 + 3389.12 + 1800.0) - 9361.68 =$$

$$= 9437.71 - 9361.68 = 76.027; \quad d.f._{\text{czynnik A}} = a - 1 = 2$$

$$SS_{\text{czynnik B}} = \sum_{j=1}^{b}\frac{\left(\sum_{i=1}^{a}\sum_{l=1}^{n_{ij}} X_{ijl}\right)^2}{\sum_{i=1}^{a} n_{ij}} - C = (2795.04 + 3158.14 + 3724.05) - 9361.68 =$$

$$= 9677.23 - 9361.68 = 315.55; \quad d.f._{\text{czynnik B}} = b - 1 = 3$$

$$SS_{\text{interakcji A x B}} = SS_{\text{grup}} - SS_{\text{czynnik A}} - SS_{\text{czynnik B}} = 394.41 - 76.027 - 315.55 = 2.83$$

$$d.f._{\text{A x B}} = (a - 1) \text{ x } (b - 1) = 6$$

$$SS_{\text{wewnątrzgrupowa (błędu)}} = SS_{\text{całk}} - SS_{\text{grup}} = 668.06 - 394.41 = 273.65$$

$$d.f._{\text{błędu}} = d.f._{\text{całk}} - d.f._{\text{grup}} = 71 - 11 = 60$$

zmienność	*SS*	*d.f.*	*MS*
całkowita	668.05944	71	
grupowa (*komórki macierzy*)	394.40806	11	35.85528
czynnik A	76.027037	2	38.01352
czynnik B	315.54955	3	105.1832
interakcje			
A x B	2.8314699	6	0.471912
wewnątrzgrupowa (błędu)	273.65139	60	4.560856

Dla H_0: Blokery nie różnią się pod względem hamowania reaktywności płytek w obecności ADP.

$$F = \frac{MS_{bloker}}{MS_{błędu}} = \frac{105.1832}{4.560856} = 23.062$$

$F_{0.05(1),3.60}$ = 2.76, dlatego odrzucamy H_0; ponieważ obliczona wartość F = 23.062 jest o wiele wyższa od tablicowej, możemy to zrobić z bardzo wysokim prawdopodobieństwem: $p < 0.0005$.

Wniosek

Blokery różnią się istotnie pod względem skuteczności hamowania reaktywności płytek w obecności ADP.

Dla H_0: nie występują różnice w reaktywności płytek krwi między nosicielami różnych genotypów polimorfizmu $Pl^{A1/A2}$.

$$F = \frac{MS_{polimorfizm}}{MS_{błędu}} = \frac{38.01352}{4.560856} = 8.335$$

$F_{0.05(1),2.60}$ = 3.15, dlatego odrzucamy H_0; ponieważ obliczona wartość F = 8.335 jest o wiele wyższa od tablicowej, możemy to zrobić z prawdopodobieństwem $p < 0.001$.

Wniosek

Występują istotne różnice w reaktywności płytek krwi między nosicielami różnych genotypów polimorfizmu $Pl^{A1/A2}$.

Dla H_0: nie ma interakcji między rodzajem blokera a rodzajem genotypu.

$$F = \frac{MS_{AxB}}{MS_{błędu}} = \frac{0.471912}{4.560856} = 0.10347$$

$F_{0.05(1),6.60}$ = 2.25, dlatego nie odrzucamy H_0; $p > 0.50$

Wniosek

Nie ma istotnej interakcji między rodzajem blokera a rodzajem genotypu w warunkach doświadczenia.

Przykład 36

Dla danych z poprzedniego przykładu zastosuj model hierarchiczny ANOVA.

Zauważmy, że pierwszy czynnik międzygrupowy (polimorfizm) jest zawsze nadrzędny w stosunku do czynnika drugiego (bloker) – każdemu dawcy krwi możemy przypisać

jeden z trzech wariatów polimorfizmu (*efekt stały*), natomiast liczba poziomów czynnika drugiego (blokera) zależy od badacza, jest manipulowalna (*efekt losowy*). U każdego nosiciela reprezentującego dowolny wariant polimorfizmu decydujemy się zbadać wpływ 4 różnych blokerów. W takim układzie poziomy 1, 2, 3 i 4 czynnika drugiego (bloker 1, bloker 2, bloker 3 i bloker 4) pojawiają się w obrębie pierwszego, drugiego i trzeciego poziomu pierwszego czynnika (polimorfizm $Pl^{A1/A1}$, $Pl^{A1/A2}$ i $Pl^{A2/A2}$). Tak więc drugi czynnik (bloker) jest zagnieżdżony w obrębie pierwszego czynnika (polimorfizm).

Dla takiego modelu hierarchicznego mamy:

zmienność	*SS*	*d.f.*	*MS*	$F = MS_{efektu}/MS_{reszt}$
czynnik A	266.581	2	133.2905	29.22488
czynnik B	78.8589	9	8.7621	1.921152
wewnątrzgrupowa (*błędu*)	273.6514	60	4.561	

Wniosek

Na podstawie wartości statystyki *F* wnioskujemy, że występują istotne różnice w reaktywności płytek krwi między nosicielami różnych genotypów polimorfizmu $Pl^{A1/A2}$ ($p< 0.0001$), ale nie między różnymi stosowanymi blokerami.

Widzimy zatem, że zastosowanie modelu hierarchicznego zmienia wynik testu w stosunku do układu, gdzie oba czynniki (polimorfizm i bloker) traktowaliśmy równorzędnie.

Przykład 37

Określano stężenie hemoglobiny we krwi pełnej u noworodków normo- i hipertroficznych w zależności od rodzaju porodu: na drodze porodu naturalnego (SN), porodu kleszczowego lub z cięciem cesarskim. Wyniki (g/100 ml) przedstawiono w tabeli poniżej.

Postawione poniżej hipotezy postanowiono zweryfikować przy poziomie istotności $\alpha = 0.05$:

- H_0: Stężenie Hb we krwi pełnej w grupach noworodków normotroficznych i hipertroficznych nie różni się, czyli $\mu_1 = \mu_2$,
- H_A: Stężenie Hb we krwi pełnej w grupach noworodków normotroficznych i hipertroficznych różni się, czyli $\mu_1 \neq \mu_2$,

- H_0: Stężenie Hb we krwi pełnej nie różni się w zależności od rodzaju porodu, czyli $\mu_1 = \mu_2 = \mu_3$,
- H_A: Stężenie Hb we krwi pełnej różni się w zależności od rodzaju porodu, czyli nie jest prawdziwe równanie $\mu_1 = \mu_2 = \mu_3$,

- H_0: nie ma interakcji między rodzajem porodu a występowaniem hipertrofii,
- H_A: występuje interakcja między rodzajem porodu a występowaniem hipertrofii.

Obliczenia:

czynnik 1	**czynnik 2**						**czynnik 1: sumy i średnie**
	SN	x_1^2	**kleszczowy**	x_2^2	**cesarskie**	x_3^2	
noworodki hipertroficzne	13.0	168.1	11.8	138.3	13.0	169.7	
	12.5	156.1	11.2	124.7	12.9	165.6	
	12.5	155.1	13.2	175.3	14.1	197.9	
	12.4	152.6	13.1	170.6	11.3	127.6	
	12.6	159.2		13.4	178.9		
	12.8	163.7		13.7	187.3		
	11.3	128.4					
	12.1	146.7					
$\sum x$	99.116544		49.22504		78.322179	$\sum x$	226.6638
$\bar{x}$	12.389568		12.30626		13.053697	$\bar{x}$ 12.59243	
$\sum x^2$	1229.775		608.8143		1027.056	$\sum x^2$	2865.646
n	8		4		6	n	18
$(\sum x)^2 / n$	1228.011		605.7761		1022.394	$(\sum x)^2 / n$	2856.181
noworodki normotroficzne	11.0	121.7	13.9	193.2	14.7	216.6	
	12.8	165.0	12.8	164.2	12.7	161.9	
	13.2	174.9	13.3	176.4	13.4	178.3	
	13.0	167.8	14.1	199.0	14.7	216.6	
	13.5	183.3	14.9	220.6	12.3	150.7	
	11.2	126.3	14.2	202.3	14.0	196.4	
	12.8	162.9		12.0	143.1		
	13.1	172.6		13.6	183.7		
	13.8	190.0		12.6	157.9		
	12.5	156.1					
	14.4	208.2					
	13.2	174.3					
$\sum x$	154.64531		83.1787		119.88324	$\sum x$	357.7072
$\bar{x}$	12.887109		13.86312		13.32036	$\bar{x}$ 13.24842	
$\sum x^2$	2003.146		1155.719		1605.198	$\sum x^2$	4764.062
n	12		6		9	n	27
$(\sum x)^2 / n$	1992.931		1153.116		1596.888	$(\sum x)^2 / n$	4742.935
$\sum_{i=1}^{a}\sum_{l=1}^{n} x$	253.76		132.40		198.21	$\sum x_{całk}$	**584.37**
n_{trofia}	20		10		15	N	45
$\overline{X}$	12.69		13.24		13.21	$\overline{X}_{całk}$	**12.98602**
$(\sum X)^2 / n$	3219.7539		1753.075		2619.0258	$(\sum X)^2 / N$	**7591.855**

$$\sum_{i=1}^{2}\sum_{j=1}^{3}\sum_{l=1}^{n} X_{ijl} = 584.371$$

$$\sum_{i=1}^{2}\sum_{j=1}^{3}\sum_{l=1}^{n} X_{ijl}^2 = 7629.708$$

a = liczba grup noworodków = 2
b = liczba rodzajów porodu = 3
N = całkowita liczba obserwacji = 45
suma całkowita dla noworodków hipertroficznych =

$$\sum_{j=1}^{3}\sum_{l=1}^{n} X_{1jl} = 226.664$$

suma całkowita dla noworodków normotroficznych =

$$\sum_{j=1}^{3}\sum_{l=1}^{n} X_{1jl} = 357.707$$

suma całkowita dla porodu SN =

$$\sum_{i=1}^{3}\sum_{l=1}^{n} X_{i1l} = 253.7619$$

suma całkowita dla porodu kleszczowego =

$$\sum_{i=1}^{3}\sum_{l=1}^{n} X_{i2l} = 132.404$$

suma całkowita dla porodu z cięciem cesarskim =

$$\sum_{i=1}^{3}\sum_{l=1}^{n} X_{i3l} = 198.21$$

$$C = \frac{\left(\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{l=1}^{n_{ij}} X_{ijl}\right)^2}{N} = 7588.6548$$

$$SS_{\text{całk}} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{l=1}^{n_{ij}} X_{ijl}^2 - C = 41.053149; \qquad d.f._{\text{całk}} = N - 1 = 44$$

$$SS_{\text{grup}} = \sum_{i=1}^{a}\sum_{j=1}^{b}\frac{\left(\sum_{l=1}^{n_{ij}} X_{ijl}\right)^2}{n_{ij}} - C = 10.46105; \qquad d.f._{\text{grup}} = ab - 1 = (3)(4) - 1 = 5$$

$$SS_{\text{czynnik A}} = \sum_{i=1}^{a}\frac{\left(\sum_{j=1}^{b}\sum_{l=1}^{n_{ij}} X_{ijl}\right)^2}{\sum_{j=1}^{b} n_{ij}} - C = 3.1997056; \qquad d.f._{\text{czynnik A}} = a - 1 = 1$$

$$SS_{\text{czynnik B}} = \sum_{j=1}^{b} \frac{\left(\sum_{i=1}^{a} \sum_{l=1}^{n_{ij}} X_{ijl} \right)^2}{\sum_{i=1}^{a} n_{ij}} - C = 4.6474182; \qquad d.f._{\text{czynnik B}} = b - 1 = 2$$

$$SS_{\text{interakcji A x B}} = SS_{\text{grup}} - SS_{\text{czynnik A}} - SS_{\text{czynnik B}} = 2.613926; \qquad d.f._{\text{A x B}} = (a - 1) \text{ x } (b - 1) = 2$$

$$SS_{\text{wewnątrzgrupowa (błędu)}} = SS_{\text{całk}} - SS_{\text{grup}} = 30.592099; \qquad d.f._{\text{błędu}} = d.f._{\text{całk}} - d.f._{\text{grup}} = 39$$

zmienność	*SS*	*d.f.*	*MS*
całkowita	41.053149	44	
grupowa (komórki)	10.461050	5	2.092210
czynnik A	3.1997056	1	3.199706
czynnik B	4.6474182	2	2.323709
interakcje			
A x B	2.613926	2	1.306963
wewnątrzgrupowa (błędu)	30.592099	39	0.784413

Dla H_0: Stężenie Hb we krwi pełnej w grupach noworodków normotroficznych i hipertroficznych nie różni się.

$$F = \frac{MS_{trofia}}{MS_{błędu}} = \frac{3.199706}{0.784413} = 4.07911$$

$F_{0.05(1),1.39} = 4.08$, dlatego nie odrzucamy H_0.

Wniosek

Stężenie Hb we krwi pełnej w grupach noworodków normotroficznych i hipertroficznych nie różni się istotnie.

Dla H_0: Stężenie Hb we krwi pełnej nie różni się w zależności od rodzaju porodu.

$$F = \frac{MS_{poród}}{MS_{błędu}} = \frac{2.32371}{0.784413} = 2.962$$

$F_{0.05(1),2.39} = 3.23$, dlatego nie odrzucamy H_0.

Wniosek

Stężenie Hb we krwi pełnej nie różni się istotnie w zależności od rodzaju porodu.

Dla H_0: nie ma interakcji między rodzajem porodu a występowaniem hipertrofii.

$$F = \frac{MS_{AxB}}{MS_{błędu}} = \frac{1.306963}{0.784413} = 1.6661673$$

$F_{0.05(1),2.39} = 3.23$, dlatego nie odrzucamy H_0.

Wniosek

Nie ma istotnej interakcji między rodzajem porodu a występowaniem hipertrofii.

Dwuczynnikowa analiza wariancji z pojedynczymi pomiarami w grupie (model z równą liczebnością grup bez powtórzeń)

Przykład 38

Określano stężenie glukozy we krwi pięcioma metodami. Oznaczenia przeprowadzone we krwi pobranej od 10 losowo dobranych ochotników w przedziale wieku 18-65 lat (bez uprzedniego diagnozowania u każdego z nich tolerancji glukozy) przedstawiono w tabeli poniżej (mg/100 ml).

Postawione poniżej hipotezy postanowiono zweryfikować przy poziomie istotności $\alpha = 0.05$:

– H_0: Oznaczone stężenie glukozy nie zależy od zastosowanej metody, czyli $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$,
– H_A: Oznaczone stężenie glukozy zależy od zastosowanej metody, czyli nie jest prawdziwe równanie $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$.

– H_0: Oznaczone stężenie glukozy nie zależy od zmienności osobniczej w badanej grupie osób,
– H_A: Na oznaczone stężenie glukozy ma wpływ zmienność osobnicza w badanej grupie osób.

Obliczenia:

	czynnik 2						
czynnik 1	**metoda 1**	**metoda 2**	**metoda 3**	**metoda 4**	**metoda 5**	$\sum x$	$(\sum X)^2$
osobnik 1	95	93	108	93	64	453	205209
2	112	103	104	105	149	573	328329
3	101	94	118	90	72	475	225625
4	104	95	91	92	78	460	211600
5	105	114	127	98	95	539	290521
6	103	99	121	96	99	518	268324
7	110	85	118	111	115	539	290521
8	114	97	115	110	91	527	277729
9	120	124	113	99	91	547	299209
10	104	117	172	112	104	609	370881

$\sum_{i=1}^{a}\sum_{l=1}^{n} x$	1068	1021	1187	1006	958	$\sum x_{całk}$	5240
$\overline{X}$	106.8	102.1	118.7	100.6	95.8	$\overline{x}_{całk}$	104.8
$\sum_{i=1}^{a}\sum_{l=1}^{n} x^2$	114532	105615	144957	101824	96994	$\sum x^2_{całk}$	563922
n_{metoda}	10	10	10	10	10	N	50
$(\sum X)^2$	1140624	1042441	1408969	1012036	917764	$\left(\sum x^2_{całk}\right)^2$	5521834

$$\sum_{i=1}^{10}\sum_{j=1}^{5}\sum_{l=1}^{n} X_{ijl} = 5240$$

$$\sum_{i=1}^{10}\sum_{j=1}^{5}\sum_{l=1}^{n} X^2_{ijl} = 563922$$

a = liczba przebadanych osób = 10
b = liczba metod = 5
N = całkowita liczba obserwacji = 50

$$C = \frac{\left(\sum_{i=1}^{a}\sum_{j=1}^{b} X_{ij}\right)^2}{N} = 549152$$

$$SS_{\text{całk}} = \sum_{i=1}^{a}\sum_{j=1}^{b} X^2_{ij} - C = 14770; \qquad d.f._{\text{całk}} = N - 1 = 49$$

$$SS_{\text{czynnik A}} = \frac{\sum_{i=1}^{a}\left(\sum_{j=1}^{b} X_{ij}\right)^2}{b} - C = 4437.6; \qquad d.f._{\text{czynnik A}} = a - 1 = 9$$

$$SS_{\text{czynnik B}} = \frac{\sum_{j=1}^{b}\left(\sum_{i=1}^{a} X_{ij}\right)^2}{a} - C = 3031.4; \qquad d.f._{\text{czynnik B}} = b - 1 = 4$$

$$SS_{\text{resztowa (interakcji A x B)}} = SS_{\text{całk}} - SS_{\text{A}} - SS_{\text{B}} = 7301; \qquad d.f._{\text{A x B}} = (a - 1) \text{ x } (b - 1) = 36$$

zmienność	*SS*	*d.f.*	*MS*
całkowita	14770	49	
osobnik	4437.6	9	493.0667
metoda	3031.4	4	757.85
resztowa (interakcje)	7301	36	202.8056

Dla H_0: Oznaczone stężenie glukozy nie zależy od zastosowanej metody, czyli $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$.

$$F = \frac{MS_{metoda}}{MS_{resztowa}} = \frac{757.85}{202.8056} = 3.7368$$

$F_{0.05(1),4.36} = 2.63$, dlatego odrzucamy H_0.

Wniosek

Oznaczone stężenie glukozy zależy od zastosowanej metody.

Dla H_0: Oznaczone stężenie glukozy nie zależy od zmienności osobniczej w badanej grupie osób.

$$F = \frac{MS_{osobnik}}{MS_{resztowa}} = \frac{493.0667}{202.8056} = 2.4312$$

$F_{0.05(1),9.36} = 2.15$, dlatego odrzucamy H_0.

Wniosek

Na oznaczone stężenie glukozy ma wpływ zmienność osobnicza w badanej grupie osób. Może się to wydać zrozumiałe, skoro nie wykluczyliśmy występowania zmienionej tolerancji glukozy u osób włączonych do badania. Wynik ten mówi nam, że zmienność międzyosobnicza jest czynnikiem, który powinniśmy rozważyć przy porównywaniu różnych metod oznaczania glukozy.

Analiza zrandomizowanych bloków

Przykład 39

Postanowiono określić wpływ 4 diet o różnych wartościach kalorycznych na przyrost masy u myszy. Dysponujemy 5 zwierzętami. Jak przeprowadzić to doświadczenie i jak zanalizować wyniki?

Gdybyśmy nie byli ograniczeni liczbą zwierząt, moglibyśmy w sposób losowy przydzielić każdą dietę tylu myszom, ile powtórzeń chcemy uzyskać. Na przykład, jeżeli wpływ każdej diety chcemy przebadać w 5 powtórzeniach, dobieramy 20 myszy: u pięciu stosuje-

my dietę 1, u następnych pięciu dietę 2, itd. W takim modelu doświadczalnym każda mysz – i w ten sposób każda obserwacja – jest niezależna od innych. Są całkowicie zrandomizowane, to znaczy każde ze zwierząt jest w sposób całkowicie przypadkowy przypisane określonej jednej diecie.

Ponieważ dysponowano jedynie 5 zwierzętami, zaplanowano doświadczenie, w którym wpływ każdej diety testowano – w przypadkowej kolejności – u każdej z 5 myszy, oczywiście zachowując odpowiednie okresy przerwy między jedną dietą a następną. W ten sposób wpływ wszystkich 4 diet testowano jednorazowo u 4 czterech różnych zwierząt. Wyniki wpływu 4 różnych diet testowanych jednorazowo tworzą blok. Pod względem koncepcyjnym termin „blok" wyników jest rozwinięciem pojęcia „para" wyników, którym posługujemy się na przykład w teście sparowanym *t* Studenta.

Rozważmy pierwsze podejście doświadczalne, w którym dysponujemy 20 zwierzętami, u których pragniemy przetestować wpływ 4 diet, każda u pięciu niezależnych myszy trzymanych w osobnych boksach. Możemy mieć wątpliwości czy warunki środowiskowe (takie jak oświetlenie, temperatura, wilgotność, przewiewność, itp.), które potencjalnie mogłyby mieć wpływ na przyrost masy ciała u zwierząt, są identyczne w każdym boksie. Idealnym rozwiązaniem zatem byłoby przetestowanie każdej z diet u każdego ze zwierząt, ponieważ każde jest trzymane w nieco odmiennych warunkach otoczenia. Kiedy tworzymy blok 4 zwierząt trzymanych w identycznych warunkach, jedynym odróżniającym je czynnikiem jest rodzaj diety. Do analizy możemy wykorzystać model III (mieszany) dwuczynnikowej ANOVA bez powtórzeń, gdzie dieta jest czynnikiem stałym, zaś numer bloku czynnikiem losowym. Nasze hipotezy zerowe będą zakładały, że: 1) dieta nie ma wpływu na przyrost masy ciała, oraz 2) przyrost masy ciała w różnych blokach jest równy. Tak naprawdę, interesuje nas jedynie wpływ czynnika stałego na wartość badanego parametru, czyli wpływ diety. Nasze hipotezy mają postać:

– H_0: przyrost masy ciała u myszy otrzymujących różne diety jest jednakowy, czyli $\mu_{dieta1} = \mu_{dieta\ 2} = \mu_{dieta\ 3} = \mu_{dieta\ 4}$,
– H_A: przyrost masy ciała u myszy otrzymujących różne diety jest różny, czyli nie jest prawdziwe równanie $\mu_{dieta1} = \mu_{dieta\ 2} = \mu_{dieta\ 3} = \mu_{dieta\ 4}$.

Przeprowadzając doświadczenie wybrano 5 myszy, z których każda była trzymana w osobnym boksie. Zwierzęta tworzące jeden blok doświadczalny były trzymane w identycznych warunkach otoczenia (temperatura, wilgotność, oświetlenie, hałas, itd.), a każde z nich otrzymywało inną dietę. Diety przydzielano każdemu z 4 poddawanych jednorazowo doświadczeniu zwierząt w sposób losowy korzystając z tablic liczb losowych. Zanotowano następujące przyrosty masy ciała (w gramach) w blokach:

Blok 1:	dieta 3 (4.9)	dieta 4 (8.8)	dieta 1 (7.0)	dieta 2 (5.3)
Blok 2:	dieta 1 (9.9)	dieta 3 (7.6)	dieta 2 (5.7)	dieta 4 (8.9)
Blok 3:	dieta 3 (5.5)	dieta 2 (4.7)	dieta 4 (8.1)	dieta 1 (8.5)
Blok 4:	dieta 4 (3.3)	dieta 2 (3.5)	dieta 1 (5.1)	dieta 3 (2.8)
Blok 5:	dieta 1 (10.3)	dieta 4 (9.1)	dieta 3 (8.4)	dieta 2 (7.7)

Obliczenia:

czynnik 2	**czynnik 1**					
blok	**dieta 1**	**dieta 2**	**dieta 3**	**dieta 4**	$B_j = \sum_{i=1}^{a} X_{ij}$	$B_j^2 = (\sum_{i=1}^{a} X_{ij})^2$
1	7.0	5.3	4.9	8.8	26.0	676.00
2	9.9	5.7	7.6	8.9	32.1	1030.41
3	8.5	4.7	5.5	8.1	26.8	718.24
4	5.1	3.5	2.8	3.3	14.7	216.09
5	10.3	7.7	8.4	9.1	35.5	1260.25
$G_i = \sum_{j=1}^{b} X_{ij}$	40.8	26.9	29.2	38.2	$\sum_{i=1}^{a}\sum_{j=1}^{b} X_{ij} = 135.1$	$\sum_{i=1}^{a}\sum_{j=1}^{b} X_{ij}^2 = 1011.95$
$\overline{X}_{ij}$	8.16	5.38	5.84	7.64		
n_{ij}	5	5	5	5	$n_a = 4$	
$G_i^2 = (\sum_{j=1}^{b} X_{ij})^2$	1664.64	723.61	852.64	1459.24	$\sum_{i=1}^{a} G_i^2 = 4700.13$	$\sum_{i=1}^{a} B_j^2 = 3900.99$

$a = 4,\ \ b = 5,\ N = ab = 20$

$$\sum_{i=1}^{a}\sum_{j=1}^{b} X_{ij}^2 = 1011.95$$

$$C = \frac{\left(\sum_{i=1}^{a}\sum_{j=1}^{b} X_{ij}\right)^2}{N} = \frac{(135.1)^2}{20} = 912.601$$

$$SS_{\text{całk}} = \sum_{i=1}^{a}\sum_{j=1}^{b} X_{ij}^2 - C = 1011.95 - 912.601 = 99.3495$$

$$SS_{\text{diet}} = \frac{\sum_{i=1}^{a} G_i^2}{b} - C = \frac{(40.8)^2 + (26.9)^2 + (29.2)^2 + (38.2)^2}{5} - 912.601 =$$

$$= \frac{1664.64 + 723.61 + 852.64 + 1459.24}{5} - 912.601 = 27.4255$$

$$SS_{\text{bloków}} = \frac{\sum_{j=1}^{b} B_j^2}{a} - C = \frac{(26.0)^2 + (32.1)^2 + (26.8)^2 + (14.7)^2 + (35.5)^2}{4} - 912.601 =$$

$$= \frac{676 + 1030.41 + 718.24 + 216.09 + 1260.25}{4} - 912.601 = 62.647$$

$$SS_{\text{resztowa}} = SS_{\text{całk}} - SS_{\text{diet}} - SS_{\text{bloków}} = 99.3495 - 27.4255 - 62.647 = 9.277$$

zmienność	*SS*	*d.f.*	*MS*
całkowita	99.3495	19	
dieta	27.4255	3	9.1418
blok	62.6470	4	15.6618
resztowa (interakcje)	9.2770	12	0.7731

Dla H_0: przyrost masy ciała u myszy otrzymujących różne diety jest jednakowy.

$$F = \frac{MS_{diet}}{MS_{resztowa}} = \frac{9.14183}{0.77308} = 11.825$$

$F_{0.05(1),3.12} = 3.49$, dlatego odrzucamy H_0; możemy to zrobić nawet przy poziomie istotności $\alpha = 0.001$.

Wniosek

Przyrost masy ciała u myszy zależy od zastosowanej diety.

Przykład 40

U 6 zdrowych ochotników zbadano wpływ dwóch blokerów na tworzenie agregatów płytkowych w modelu dynamicznym (mieszadło rotacyjne): antagonisty receptora dla fibrynogenu płytek krwi – GR144053F oraz antagonisty czynnika von Willebranda – kwasu aurinotrikarboksylowego (ATA). Frakcje agregatów płytkowych oznaczano na podstawie rozpraszania światła w cytometrze przepływowym. Doświadczenie przeprowadzono w ten sposób, że krew pobraną od każdego ochotnika rozdzielano na trzy równe objętości oraz do każdej porcji dodawano albo GR144053F, albo ATA, albo odpowiadającą objętość soli fizjologicznej (kontrola). Wszystkie próby inkubowano w takich samych warunkach, następnie znakowano przeciwciałami monoklonalnymi, utrwalano i przeprowadzano pomiar cytometryczny. Na podstawie uzyskanych wyników należy zweryfikować parę hipotez:

– H_0: wielkość frakcji agregatów płytkowych jest taka sama w nieobecności oraz w obecności testowanych antagonistów, czyli $\mu_{\text{kontrola}} = \mu_{\text{GR144053F}} = \mu_{\text{ATA}}$,
– H_A: wielkość frakcji agregatów płytkowych zmienia się pod wpływem testowanych antagonistów, czyli nie jest prawdziwe równanie $\mu_{\text{kontrola}} = \mu_{\text{GR144053F}} = \mu_{\text{ATA}}$.

Obliczenia:

czynnik 2	**czynnik 1**				
blok	**bez blokera**	**GR144053F**	**ATA**	$B_j = \sum_{i=1}^{a} X_{ij}$	$B_j^2 = (\sum_{i=1}^{a} X_{ij})^2$
1	1.0	4.3	1.4	6.7	44.89
2	2.4	15.6	5.8	23.8	566.44
3	7.1	10.3	2.5	19.9	396.01
4	5.1	2.8	5.1	13.0	169.00
5	2.2	1.5	3.7	7.4	54.76
6	15.9	10.9	2.3	29.1	846.81
$G_i = \sum_{j=1}^{b} X_{ij}$	33.7	45.4	20.8	$\sum_{i=1}^{a}\sum_{j=1}^{b} X_{ij} = 99.9$	$\sum_{i=1}^{a}\sum_{j=1}^{b} X_{ij}^2 = 924.51$
$\overline{X}_{ij}$	5.62	7.57	3.47		
n_{ij}	6	6	6	$n_a = 3$	
$G_i^2 = (\sum_{j=1}^{b} X_{ij})^2$	1135.69	2061.16	432.64	$\sum_{i=1}^{a} G_i^2 = 3629.49$	$\sum_{i=1}^{a} B_j^2 = 2077.91$

$a = 3, \quad b = 6, \quad N = ab = 18$

$$\sum_{i=1}^{a}\sum_{j=1}^{b} X_{ij}^2 = 924.51$$

$$C = \frac{\left(\sum_{i=1}^{a}\sum_{j=1}^{b} X_{ij}\right)^2}{N} = \frac{(99.9)^2}{18} = 554.445$$

$$SS_{\text{całk}} = \sum_{i=1}^{a}\sum_{j=1}^{b} X_{ij}^2 - C = 924.51 - 554.445 = 370.065$$

$$SS_{\text{bloker}} = \frac{\sum_{i=1}^{a} G_i^2}{b} - C = 50.47$$

$$SS_{\text{bloków}} = \frac{\sum_{j=1}^{b} B_j^2}{a} - C = 138.19167$$

$$SS_{\text{resztowa}} = SS_{\text{całk}} - SS_{\text{blokerów}} - SS_{\text{bloków}} = 370.065 - 50.47 - 138.19167 = 181.403$$

zmienność	*SS*	*d.f.*	*MS*
całkowita	370.06500	17	
bloker	50.47000	2	25.23500
blok	138.19167	5	27.63833
resztowa (interakcje)	181.40300	10	18.140300

Dla H_0: wielkość frakcji agregatów płytkowych jest taka sama w nieobecności oraz w obecności testowanych antagonistów.

$$F = \frac{MS_{bloker}}{MS_{resztowa}} = \frac{25.235}{18.1403} = 1.3911$$

$F_{0.05(1),2.10} = 4.10$, dlatego nie odrzucamy H_0.

Wniosek

Wielkość frakcji agregatów płytkowych nie zmienia się pod wpływem testowanych antagonistów.

Testy porównań wielokrotnych

Przykład 41

W tabeli poniżej przedstawiono stężenia HDL (mg/100 ml) w surowicy krwi u mężczyzn w wieku 35-60 lat zamieszkujących 5 alpejskich wiosek. Należy obliczyć, w której grupie stężenie HDL jest istotnie najwyższe.

wioska 1	wioska 2	wioska 3	wioska 4	wioska 5
28.2	39.6	46.3	41.0	56.3
33.2	40.8	42.1	44.1	54.1
36.4	37.9	43.5	46.4	59.4
34.6	37.1	48.8	40.2	62.7
29.1	43.6	43.7	38.6	60.0
31.0	42.4	40.1	36.3	57.3
$\overline{X}_1 = 32.1$	$\overline{X}_2 = 40.2$	$\overline{X}_3 = 44.1$	$\overline{X}_4 = 41.1$	$\overline{X}_5 = 58.3$
$n_1 = 6$	$n_2 = 6$	$n_3 = 6$	$n_4 = 6$	$n_5 = 6$

Przeprowadzamy analizę wariancji dla pary hipotez:
H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$,
H_A: hipoteza zerowa jest fałszywa.

zmienność	*SS*	*d.f.*	*MS*
całkowita	2437.572	29	
międzygrupowa	2193.442	4	548.3605
wewnątrzgrupowa	244.130	25	9.7652

Wartość *F* wynosi

$$\frac{MS_{międzygrupowa}}{MS_{wewnątrzgrupowa}} = \frac{548.3605}{9.7652} = 56.15$$

Ponieważ tablicowa wartość $F_{0.0014,25}$ = 6.49 jest mniejsza od obliczonej, odrzucamy hipotezę zerową z prawdopodobieństwem ponad 99.9%.

Na podstawie analizy wariancji stwierdziliśmy, że średnie nie są sobie równe, toteż możemy zastosować test porównań wielokrotnych Tukeya, aby wykazać czy występują istotne różnice między poszczególnymi grupami. Średnie w grupach porządkujemy w szeregu rosnącym:

1	2	4	3	5
32.1	40.2	41.1	44.1	58.3

Testujemy każdą z par średnich na podstawie zerowych hipotez indywidualnych postaci:

H_0: $\mu_B = \mu_A$

Analogicznie jak przy liczeniu błędu standardowego do statystyki testu *t* Studenta, nasz błąd standardowy różnicy między średnimi w każdej parze jest pochodną wewnątrzgrupowego błędu średniokwadratowego ($MS_{błędu}$) i wynosi:

$$SE = \sqrt{\frac{s^2}{n}} = \sqrt{\frac{(MS_{błędu})}{n}} = \sqrt{\frac{9.7652}{6}} = \sqrt{1.6275} = 1.28$$

Podobnie, *per analogiam* do statystyki testu *t* Studenta, wartość *q* statystyki testu Tukeya dla każdej pary średnich jest obliczana jako:

$$q = \frac{\overline{X}_B - \overline{X}_A}{SE}$$

Jeżeli obliczona wartość *q* jest równa lub większa od *q* odczytanego w tablicach (zależnego od α, *d.f.* = $N - k$ dla $MS_{błędu}$ w analizie wariancji i liczby porównywanych grup, *k*), to odrzucamy hipotezę zerową H_0: $\mu_B = \mu_A$.

Obliczenia statystyki testu Tukeya przedstawia tabela:

porównanie (*B vs. A*)	**różnica** ($\overline{X}_A - \overline{X}_B$)	**SE**	*q*	$q_{005.24.5}$	**decyzja**
5 *vs.* 1	58.3 – 32.1 = 26.2	1.28	20.47	4.166	odrzucić H_0: $\mu_5 = \mu_1$
5 *vs.* 2	58.3 – 40.2 = 18.1	1.28	14.14	4.166	odrzucić H_0: $\mu_5 = \mu_2$
5 *vs.* 4	58.3 – 41.1 = 17.2	1.28	13.44	4.166	odrzucić H_0: $\mu_5 = \mu_4$
5 *vs.* 3	58.3 – 44.1 = 14.2	1.28	11.09	4.166	odrzucić H_0: $\mu_5 = \mu_3$
3 *vs.* 1	44.1 – 32.1 = 12.0	1.28	9.38	4.166	odrzucić H_0: $\mu_3 = \mu_1$
3 *vs.* 2	44.1 – 40.2 = 3.9	1.28	3.05	4.166	przyjąć H_0: $\mu_3 = \mu_2$
3 *vs.* 4	nie testujemy				
4 *vs.* 1	44.1 – 32.1 = 9.0	1.28	7.03	4.166	odrzucić H_0: $\mu_4 = \mu_1$
4 *vs.* 2	nie testujemy				
2 *vs.* 1	40.2 – 32.1 = 8.1	1.28	6.33	4.166	odrzucić H_0: $\mu_2 = \mu_1$

Zapis w tabeli może wzbudzić wątpliwość u uważnych Czytelników, dlaczego nie testujemy porównań grup 3 *vs*. 4 oraz 4 *vs*. 2. Otóż, dla porównania średnich 3 *vs*. 2 różnica między średnimi była na tyle mała, że obliczona statystyka q testu Tukeya nie pozwoliła nam na odrzucenie hipotezy zerowej. Skoro tak, to dla różnic jeszcze mniejszych (3 *vs*. 4: 44.1 – 41.1 = 3.0 lub 4 *vs*. 2 : 41.1 – 40.2 = 0.9), tym bardziej nie będziemy mieli podstaw do odrzucenia hipotezy zerowej. Zatem nie mamy podstaw do testowania istotności tych różnic.

Wniosek

Ponieważ $\mu_1 \neq \mu_2 = \mu_4 = \mu_3 \neq \mu_5$ możemy stwierdzić, że stężenie HDL jest najwyższe wśród mieszkańców wioski 5, a najniższe wśród mieszkańców z wioski 1.

Przykład 42

U pacjentów, którzy zostali poddani operacji kardiochirurgicznej, stosowano po zabiegu leki przeciwpłytkowe w celu zmniejszenia ryzyka pooperacyjnych okluzji naczyniowych. Sprawność hemostatyczną płytek krwi oceniano przy użyciu analizatora funkcji płytek PFA-100™ i wyrażano jako czas okluzji (sek.) w kasetach z kolagenem i ADP (CADP) (krótszy czas okluzji oznacza większą reaktywność płytek krwi):

	lek 1	**lek 2**	**lek 3**	**lek 4**	
	61	69	103	88	
	57	68	102	84	
	65	74	100	83	
	59	66	97	86	
	62	70		90	
n_i	5	5	4	5	$N = \sum_{i=1}^{k} n_i = 19$
$\sum_{j=1}^{n_i} X_{ij}$	304	347	402	431	$\sum_{i=1}^{k}\sum_{j=1}^{n} X_{ij} = 1484$
$\overline{X}_i$	60.8	69.4	100.5	86.2	$\sum_{i=1}^{k}\sum_{j=1}^{n} X_{ij}^2 = 120244$
$\frac{\left(\sum_{j=1}^{n_i} X_{ij}\right)^2}{n_i}$	18483.2	24081.8	40401	37152.2	$\sum_{i=1}^{k} \frac{\left(\sum_{j=1}^{n} X_{ij}\right)^2}{n_i} = 120118.2$

Nasze hipotezy:
H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$
H_A: H_0 nie jest prawdziwa
weryfikujemy przy użyciu jednoczynnikowej analizy wariancji (model ANOVA I).

Obliczenia:

$\sum_{i=1}^{k}\sum_{j=1}^{n} X_{ij} = 1484$; $\quad d.f._{\text{całk}} = N - 1 = 19 - 1 = 18$

$$\sum_{i=1}^{k}\sum_{j=1}^{n} X_{ij}^2 = 120244; \qquad d.f._{\text{grup}} = k - 1 = 4 - 1 = 3$$

$$\sum_{i=1}^{k} \frac{\left(\sum_{j=1}^{n} X_{ij}\right)^2}{n_i} = 120118.2; \qquad d.f._{\text{błędu}} = N - k = 19 - 4 = 15$$

$$C = \frac{\left(\sum_{i=1}^{k}\sum_{j=1}^{n} X_{ij}\right)^2}{N} = \frac{(1484)^2}{19} = 115908.2$$

$$SS_{\text{całk}} = \sum_{i=1}^{k}\sum_{j=1}^{n} X_{ij}^2 - C = 120244 - 115908.2 = 4335.8$$

$$SS_{\text{grup}} = \sum_{i=1}^{k} \frac{\left(\sum_{j=1}^{n} X_{ij}\right)^2}{n_i} - C = 120118.2 - 115908.2 = 4209.99$$

$$SS_{\text{błędu}} = SS_{\text{całk}} - SS_{\text{grup}} = 4335.8 - 4209.99 = 125.8$$

zmienność	*SS*	*d.f.*	*MS*
całkowita	4335.789	18	
grup	4209.989	3	1403.330000
błędu	125.800	15	8.386667

$$F = \frac{MS_{grup}}{MS_{błędu}} = \frac{1403.33}{8.38667} = 167.33$$

$F_{0.05(1),3.15}$ = 3.29, dlatego odrzucamy H_0; ponieważ $F_{dośw}$ jest o wiele większe od $F_{0.05(1),3.15}$ = 3.29, możemy na poziomie istotności mniejszym niż $\alpha = 0.0005$ wnioskować, że wpływ badanych leków na czas okluzji jest istotnie różny.

Aby dowiedzieć się, który z leków działa najskuteczniej (kiedy czas okluzji jest najwyższy), stosujemy test porównań wielokrotnych Tukeya.

Ponieważ grupy badane nie są jednakowo liczne, błąd standardowy SE obliczamy według równania:

$$SE = \sqrt{\frac{s^2}{2}\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}$$

Średnie w grupach porządkujemy w szeregu rosnącym:

grupa	**1**	**2**	**4**	**3**
średnia	60.8	69.4	86.2	100. 5
liczebność	5	5	5	4

Testujemy każdą z par średnich na podstawie zerowych hipotez indywidualnych postaci:
H_0: $\mu_B = \mu_A$
$k = 4$
$s^2 = MS_{błędu} = 8.3867$
$d.f._{błędu} = 15$

porównanie (*B vs. A*)	**różnica** ($\overline{X}_A - \overline{X}_B$)	**SE***	q	$q_{0.05.15.4}$	**decyzja**
3 *vs.* 1	100.5 – 60.8 = 39.7	1.373681	28.90045	4.076	odrzucić H_0: $\mu_3 = \mu_1$
3 *vs.* 2	100.5 – 69.4 = 31.1	1.373681	22.6399	4.076	odrzucić H_0: $\mu_3 = \mu_2$
3 *vs.* 4	100.5 – 86.2 = 14.3	1.373681	10.40998	4.076	odrzucić H_0: $\mu_3 = \mu_4$
4 *vs.* 1	86.2 – 60.8 = 25.4	1.295119	19.6121	4.076	odrzucić H_0: $\mu_4 = \mu_1$
4 *vs.* 2	86.2 – 69.4 = 16.8	1.295119	12.97178	4.076	odrzucić H_0: $\mu_4 = \mu_2$
2 *vs.* 1	69.4 – 60.8 = 8.6	1.295119	6.640316	4.076	odrzucić H_0: $\mu_2 = \mu_1$

* ponieważ $n_3 \neq n_1, n_2\, n_4$ to $SE = \sqrt{\left(\frac{8.38667}{2}\right)\left(\frac{1}{4}+\frac{1}{5}\right)} = \sqrt{(4.193333)(0.25+0.20)} = \sqrt{1.887} = 1.373681$

dla $n_4 = n_1 = n_2$ $\quad SE = \sqrt{\left(\frac{8.38667}{5}\right)} = \sqrt{1.6773} = 1.295119$

Na podstawie obliczeń wnioskujemy, że $\mu_1 \neq \mu_2 \neq \mu_4 \neq \mu_3$, czyli z prawdopodobieństwem przynajmniej 95% możemy powiedzieć, że leki charakteryzują się istotnie różną skutecznością: lek 3 jest najskuteczniejszy, zaś lek 1 najmniej skuteczny w hamowaniu reaktywności płytek krwi u badanych pacjentów.

Przykład 43

Dla danych z przykładu 41 zastosujmy test porównań wielokrotnych Newmana-Keulsa.

Uporządkowany szereg średnich w grupach wygląda następująco:

1	2	4	3	5
32.1	40.2	41.1	44.1	58.3

porównanie (*B vs. A*)	różnica $(\overline{X}_A - \overline{X}_B)$	SE	*q*	*p*	$Q_{005.24.p}$	decyzja	
5 *vs.* 1	58.3 – 32.1 = 26.2	1.28	20.47	5	4.166	odrzucić	H_0: $\mu_5 = \mu_1$
5 *vs.* 2	58.3 – 40.2 = 18.1	1.28	14.14	4	3.901	odrzucić	H_0: $\mu_5 = \mu_2$
5 *vs.* 4	58.3 – 41.1 = 17.2	1.28	13.44	3	3.532	odrzucić	H_0: $\mu_5 = \mu_4$
5 *vs.* 3	58.3 – 44.1 = 14.2	1.28	11.09	2	2.919	odrzucić	H_0: $\mu_5 = \mu_3$
3 *vs.* 1	44.1 – 32.1 = 12.0	1.28	9.38	4	3.901	odrzucić	H_0: $\mu_3 = \mu_1$
3 *vs.* 2	44.1 – 40.2 = 3.9	1.28	3.05	3	3.532	przyjąć	H_0: $\mu_3 = \mu_2$
3 *vs.* 4	nie testujemy						
4 *vs.* 1	44.1 – 32.1 = 9.0	1.28	7.03	3	3.532	odrzucić	H_0: $\mu_4 = \mu_1$
4 *vs.* 2	nie testujemy						
2 *vs.* 1	40.2 – 32.1 = 8.1	1.28	6.33	2	2.190	odrzucić	H_0: $\mu_2 = \mu_1$

Wniosek

Ponieważ $\mu_1 \neq \mu_2 = \mu_4 = \mu_3 \neq \mu_5$ możemy stwierdzić, że stężenie HDL jest najwyższe wśród mieszkańców wioski 5, a najniższe wśród mieszkańców z wioski 1.

Test Newmana-Keulsa jest pod względem matematycznym bardzo podobny do Testu Tukeya z jednym wyjątkiem: krytyczne wartości statystyki $q_{\alpha,v,p}$ zależą od rozstępu rang średnich porównywanych grup, *p*. Na przykład, porównując średnią grupy 5 i średnią grupy 1, rozstęp rang średnich wynosi $p = 5$, zaś dla porównania średnich 5 *vs*. 2 wynosi on $p = 4$, itd. Wartości krytyczne statystyki *q* odczytujemy z tych samych tablic rozkładu *q* co w przypadku posługiwania się testem Tukeya.

Wynik analizy tego testu dla danych tego przykładu jest identyczny z wynikiem analizy testu Tukeya (Przykład 41), ale nie zawsze tak jest. Ogólnie, test Newmana-Keulsa jest testem o mniejszej mocy niż test Tukeya, zatem częściej wykrywa istotność różnic między grupami (częściej wykrywa różnice fałszywie istotne) oraz częściej prowadzi do odrzucenia hipotezy zerowej w przypadkach gdy jest ona prawdziwa.

Przykład 44

W sześciu grupach pacjentów z chorobą niedokrwienną serca badano wpływ 7-dniowego stosowania dwóch dawek kwasu acetylosalicylowego (ASA) oraz trzech różnych antagonistów receptora dla fibrynogenu: Integriliny, Aggrastatu oraz Rheo-Pro (Abciximab) na indukowaną ADP agregację płytek badaną w osoczu bogatopłytkowym metodą turbidymetryczną. Na podstawie zebranych wyników należy obliczyć, czy każdy z badanych leków przyczynia się do zahamowania czynności płytek ocenianej metodą agregacji turbidymetrycznej.

kontrola	ASA 75 mg	ASA 150 mg	Integrilina	Aggrastat	Abciximab
0.386	0.267	0.328	0.343	0.145	0.153
0.426	0.142	0.242	0.320	0.185	0.083
0.405	0.240	0.255	0.242	0.191	0.109
0.309	0.269	0.236	0.242	0.148	0.077
0.306	0.106	0.107	0.128	0.178	0.178
0.244	0.146	0.243	0.272	0.101	0.070
0.329	0.274	0.188	0.179	0.140	0.066
0.295	0.236	0.185	0.353	0.080	0.144
0.316	0.141	0.111	0.117	0.078	0.073
0.423	0.243	0.141	0.168	0.136	0.128

0.162	0.102	0.057	0.072	0.123	0.051
0.314	0.215	0.126	0.131	0.141	0.132
0.276	0.137	0.190	0.254	0.202	0.099
0.386	0.238	0.160	0.257	0.040	0.138
0.301	0.189	0.157	0.229	0.151	0.117
0.229	0.171	0.291	0.297	0.129	0.104
0.231	0.144	0.055	0.108	0.101	0.084
0.325					
0.381					
0.307					
0.321					
0.361					
0.245					
0.253					
0.328					
0.257					

$\overline{X}_1 = 0.312$	$\overline{X}_2 = 0.192$	$\overline{X}_3 = 0.181$	$\overline{X}_4 = 0.218$	$\overline{X}_5 = 0.133$	$\overline{X}_6 = 0.106$
$n_1 = 26$	$n_2 = 17$	$n_3 = 17$	$n_4 = 17$	$n_5 = 17$	$n_6 = 17$

Do oceny zastosujemy test Dunnetta, który pozwala na sprawdzenie, czy średnia jednej grupy (nazwanej umownie grupą kontrolną) różni się istotnie od każdej ze średnich $k-1$ innych grup (badanych). W tym przypadku nie jesteśmy zainteresowani, czy istotne są wzajemne różnice między każdą z grup, lecz testujemy jedynie $k-1$ porównań z grupą „kontrolną". W naszym przypadku będzie to grupa pierwsza.

Jako pierwszy krok obliczeń przeprowadzamy analizę wariancji dla pary hipotez:

– H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6$
– H_A: hipoteza zerowa jest fałszywa

zmienność	*SS*	*d.f.*	*MS*
całkowita	0.996020	110	
międzygrupowa	0.565073	5	0.113015
wewnątrzgrupowa	0.430947	105	0.004104

Wartość F wynosi $\dfrac{0.113015}{0.004104} = 27.536$

Ponieważ tablicowa wartość $F_{0.001,5,104} = 5.751$ jest o wiele mniejsza od obliczonej, odrzucamy hipotezę zerową z prawdopodobieństwem co najmniej 99.99%.

Skoro na podstawie analizy wariancji stwierdziliśmy, że średnie w 6 grupach nie są sobie równe, to możemy zastosować test porównań grupy kontrolnej z każdą z grup badanych (Dunnetta), aby wykazać, czy badane leki powodują istotne różnice stopnia agregacji płytek aktywowanych ADP. W celu oszacowania wartości statystyki q testu Dunnetta, policzymy najpierw błąd standardowy dla tego testu. Błąd ten liczymy ogólnie posługując się równaniem:

$$SE = \sqrt{\frac{2s^2}{n}},$$

a w naszym konkretnym przypadku – ponieważ liczności grup nie są identyczne – zastosujemy wzór:

$$SE = \sqrt{s^2\left(\frac{1}{n_A} + \frac{1}{n_{control}}\right)},$$

gdzie:
n_A oznacza liczebność grupy badanej (w naszym przypadku grupy od 2 do 6 o liczebności po 17 obserwacji każda),
$n_{control}$ jest liczebnością grupy kontrolnej (w naszym przypadku grupa 1 o liczebności 26 obserwacji), zaś

$s^2 = MS_{błędu} = 0.004104.$

Błąd standardowy testu Dunnetta będzie wynosić:

$$SE = \sqrt{0.004104 * \left(\frac{1}{17} + \frac{1}{26}\right)} = \sqrt{0.0003884} = 0.006233.$$

Na jego podstawie możemy obliczyć wartości statystyki q testu Dunnetta w następujący sposób:

$$q = \frac{\overline{X}_{control} - \overline{X}_A}{SE}.$$

Ponieważ interesuje nas czy badane leki powodują obniżenie agregacji płytek krwi, powinniśmy sprawdzić, czy średnia dla grupy pierwszej μ_1 jest większa od każdej ze średnich pozostałych pięciu grup (μ_A). Musimy zatem przetestować hipotezę zerową: H_0: $\mu_1 \leq \mu_A$ przeciwstawną do hipotezy alternatywnej: H_A: $\mu_1 > \mu_A$.

porównanie (kontrola *vs*. A)	**różnica**	**SE**	**(q)**	$q_{0.01(1),105,6}$	**decyzja**
1 *vs*. 2	0.312 – 0.192 = 0.120	0.00623251	19.321	2.91	odrzucić H_0: $\mu_1 \leq \mu_A$
1 *vs*. 3	0.312 – 0.181 = 0.132	0.00623251	21.103	2.91	odrzucić H_0: $\mu_1 \leq \mu_A$
1 *vs*. 4	0.312 – 0.218 = 0.094	0.00623251	15.062	2.91	odrzucić H_0: $\mu_1 \leq \mu_A$
1 *vs*. 5	0.312 – 0.133 = 0.179	0.00623251	28.675	2.91	odrzucić H_0: $\mu_1 \leq \mu_A$
1 *vs*. 6	0.312 – 0.106 = 0.206	0.00623251	33.041	2.91	odrzucić H_0: $\mu_1 \leq \mu_A$

Zauważmy, że posługujemy się jednostronnym testem Dunnetta (stąd zapis: $q_{0.01(1),105,6}$, gdzie 0.01 oznacza przyjęty poziom istotności dla testu jednostronnego (1), dla liczby stopni swobody błędu $d.f._{błędu}$ = 105 oraz liczby grup k = 6), ponieważ zależy nam na przyjęciu hipotezy, mówiącej, że agregacja płytek krwi jest niższa w wyniku podawania leku, a nie różna od agregacji płytek w grupie kontrolnej.

Ponieważ dla każdego porównania średniej grupy kontrolnej z odpowiednią średnią grupy badanej znaleźliśmy, że nasze $q_{dośw} > q_{0.01(1),105,6}$, zatem w każdym przypadku możemy

odrzucić hipotezę zerową H_0: $\mu_1 \leq \mu_A$ na korzyść hipotezy alternatywnej H_A: $\mu_1 > \mu_A$, oraz wnioskować, iż każdy z leków wpływa na obniżenie agregacji płytek krwi pod wpływem ADP mierzonej metodą turbidymetryczną.

Przykład 45

Dla danych z przykładu 41 zastosujmy teraz test wielokrotnych kontrastów Scheffe'go. Przypomnijmy, że w przykładzie tym porównywaliśmy stężenia HDL w surowicy krwi wśród mężczyzn zamieszkujących 5 różnych wiosek na terenie Alp. Interesowało nas, w której wiosce mężczyźni charakteryzują się najwyższymi stężeniami HDL we krwi.

wioska 1	wioska 2	wioska 3	wioska 4	wioska 5
28.2	39.6	46.3	41.0	56.3
33.2	40.8	42.1	44.1	54.1
36.4	37.9	43.5	46.4	59.4
34.6	37.1	48.8	40.2	62.7
29.1	43.6	43.7	38.6	60.0
31.0	42.4	40.1	36.3	57.3

$\overline{X}_1 = 32.1$ $\overline{X}_2 = 40.2$ $\overline{X}_3 = 44.1$ $\overline{X}_4 = 41.1$ $\overline{X}_5 = 58.3$

$n_1 = 6$ $n_2 = 6$ $n_3 = 6$ $n_4 = 6$ $n_5 = 6$

Przeprowadzona analiza wariancji dla pary hipotez:
– H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$
– H_A: hipoteza zerowa jest fałszywa

wykazała, że

zmienność	*SS*	*d.f.*	*MS*
całkowita	437.572	29	
międzygrupowa	2193.442	4	548.3605
wewnątrzgrupowa	244.130	25	9.7652

Policzona na podstawie tych wartości błędów wartość statystyki *F* wynosiła $\frac{548.3605}{9.7652} = 56.15$ i była większa od $F_{0.001,4,25} = 6.49$ odczytanego w tablicach, co upoważniło nas do odrzucenia hipotezy zerowej i wnioskowania, że średnie stężenia HDL w różnych wioskach nie są sobie równe. Zastosowany test porównań wielokrotnych Tukeya dla szeregu średnich

1	2	4	3	5
32.1	40.2	41.1	44.1	58.3

wykazał, że $\mu_1 \neq \mu_2 = \mu_4 = \mu_3 \neq \mu_5$, a zatem mogliśmy stwierdzić, że stężenie HDL jest najwyższe wśród mieszkańców wioski 5, a najniższe wśród mieszkańców z wioski 1.

Czy wniosek będzie taki sam, gdy do opracowania tych danych zastosujemy test Scheffe'go?

Statystykę testu Scheffe'go, S, liczymy w następujący sposób:

$$S = \frac{\left|\overline{X}_B - \overline{X}_A\right|}{SE},$$

gdzie $\overline{X}_B$ oraz $\overline{X}_A$ oznaczają średnie porównywanych grup, a SE oznacza błąd standardowy tego testu i obliczany jest jako:

$$SE = \sqrt{s^2\left(\frac{1}{n_B} + \frac{1}{n_A}\right)}$$

lub – gdy liczebności obu porównywanych grup, n_A i n_B, są jednakowe (tak jak w naszym analizowanym przykładzie) – jako:

$$SE = \sqrt{\frac{2s^2}{n}}.$$

W równaniach tych wyrażenie s^2 oznacza oczywiście wewnątrzgrupowy błąd średniokwadratowy ($MS_{błędu}$). Ponieważ $MS_{błędu} = 9.7652$ oraz $n = 6$, mamy:

$$SE = \sqrt{\frac{2s^2}{n}} = \sqrt{\frac{2 * 9.7652}{6}} = \sqrt{3.255} = 1.804$$

Wartość krytyczną dla testu Scheffe'go liczymy jako:

$$S_\alpha = \sqrt{(k-1)F_{\alpha(1),k-1,N-k}} = \sqrt{(5-1) * (2.76)} = \sqrt{11.04} = 3.323,$$

gdzie k oznacza liczbę grup (5), zaś $F_{\alpha(1),k-1,N-k} = 2.76$ oznacza wartość krytyczną testu F dla poziomu istotności α (0.05), $k-1$ stopni swobody dla grup (4) oraz $N-k$ stopni swobody dla błędu (25).

porównanie (B *vs.* A)	**różnica ($\overline{X}_A - \overline{X}_B$)**	**SE**	**S**	**$S_{005(1),4,25}$**	**Decyzja**	
5 *vs.* 1	58.3 – 32.1 = 26.2	1.804	14.522	3.32265	odrzucić	H_0: $\mu_5=\mu_1$
5 *vs.* 2	58.3 – 40.2 = 18.1	1.804	10.032	3.32265	odrzucić	H_0: $\mu_5=\mu_2$
5 *vs.* 4	58.3 – 41.1 = 17.2	1.804	9.533	3.32265	odrzucić	H_0: $\mu_5=\mu_4$
5 *vs.* 3	58.3 – 44.1 = 14.2	1.804	7.871	3.32265	odrzucić	H_0: $\mu_5=\mu_3$
3 *vs.* 1	44.1 – 32.1 = 12.0	1.804	6.651	3.32265	odrzucić	H_0: $\mu_3=\mu_1$
3 *vs.* 2	44.1 – 40.2 = 3.9	1.804	2.162	3.32265	przyjąć	H_0: $\mu_3=\mu_2$
3 *vs.* 4	nie testujemy					
4 *vs.* 1	44.1 – 32.1 = 9.0	1.804	4.988	3.32265	odrzucić	H_0: $\mu_4=\mu_1$
4 *vs.* 2	nie testujemy					
2 *vs.* 1	40.2 – 32.1 = 8.1	1.804	4.490	3.32265	odrzucić	H_0: $\mu_2=\mu_1$

Wniosek

Na podstawie obliczeń wnioskujemy, że $\mu_1 \neq \mu_2 = \mu_4 = \mu_3 \neq \mu_5$, czyli możemy stwierdzić, że stężenie HDL jest najwyższe wśród mieszkańców wioski 5, a najniższe wśród mieszkańców z wioski 1.

Widzimy, że test Scheffe'go przywiódł nas do takiego samego wniosku jak test Tukeya. Nie zawsze tak jest, ponieważ test Tukeya charakteryzuje się większą mocą niż test Scheffe'go, czyli wykrywa istotne różnice z mniejszą częstością. Z uwagi na tą mniejszą moc, i w związku z tym większe prawdopodobieństwo popełnienia błędu II rodzaju, test Scheffe'go nie jest zalecany w przypadkach, gdzie możemy wykorzystać test Tukeya lub test Newmana-Keulsa. Istnieją jednak szczególne przypadki, do których test ten nadaje się o wiele lepiej niż jakakolwiek inna procedura porównań wielokrotnych. Są to tzw. „wielokrotne kontrasty", gdzie porównania średnich z kilku grup zestawia się w szczególny sposób. Na przykład, chcąc wykazać, że średnie stężenie HDL wśród wszystkich mężczyzn zamieszkujących wioski 2, 3 i 4, nie jest istotnie różne od stężenia HDL obserwowanego u mężczyzn z wioski 5, powinniśmy zapisać:

H_0: $(\mu_2 + \mu_3 + \mu_4)/3 - \mu_5 = 0$.

W takim zestawieniu zakładamy, że zależność tą można wyrazić jako:

H_0: $\mu_2/3 + \mu_3/3 + \mu_4/3 - \mu_5 = 0$,

w której średnim μ_2, μ_3, μ_4 i μ_5 przypisać można współczynniki proporcjonalności $c_2 = \frac{1}{3}$, $c_3 = \frac{1}{3}$, $c_4 = \frac{1}{3}$ oraz $c_5 = -1$. Warto zauważyć, iż suma wszystkich dobranych w ten sposób współczynników stanowi zawsze 0, dlatego że każdy z nich stanowi jakąś porcję jedności w zestawieniu średnich. Wartość statystyki S testu jest wtedy liczona jako:

$$S = \frac{\left|\sum c_i \overline{X}_i\right|}{SE},$$

gdzie:

c_i oznacza współczynnik proporcjonalności dla średniej $\overline{X}_i$, a wartość błędu standardowego testu S obliczamy jako:

$$SE = \sqrt{s^2\left(\sum \frac{c_i^2}{n_i}\right)};$$

jak wyżej, s^2 oznacza oczywiście wewnątrzgrupowy błąd średniokwadratowy ($MS_{błędu}$), zaś n_i jest liczebnością poszczególnych grup.

Zwróćmy uwagę, że wartość licznika równania statystyki S może być zarówno dodatnia jak i ujemna, stąd w równaniu występuje moduł sumy iloczynów średnich i współczynników proporcjonalności w badanym równaniu.

Dla naszego równania $(\mu_2 + \mu_3 + \mu_4)/3 - \mu_5 = 0$ mamy zatem:

$$SE = \sqrt{9.7652\left[\frac{\left(\frac{1}{3}\right)^2}{6}+\frac{\left(\frac{1}{3}\right)^2}{6}+\frac{\left(\frac{1}{3}\right)^2}{6}+\frac{(1)^2}{6}\right]} = \sqrt{9.7652\left[\left(\frac{1}{54}\right)+\left(\frac{1}{54}\right)+\left(\frac{1}{54}\right)+\left(\frac{1}{6}\right)\right]} =$$

$$= \sqrt{9.7652\left[\frac{12}{54}\right]} = \sqrt{2.17} = 1.473$$

$$\text{oraz } S = \frac{\left|\left(\frac{40.2}{3}\right)+\left(\frac{44.1}{3}\right)+\left(\frac{41.1}{3}\right)-58.3\right|}{1.473} = \frac{|13.4+14.7+13.7-58.3|}{1.473} = \frac{|-16.5|}{1.473} = 11.202$$

Dla poziomu istotności $\alpha = 0.05$, wartość krytyczna testu S,

$$S_{0.05} = \sqrt{(k-1)F_{\alpha(1),k-1,N-k}} = \sqrt{(5-1)(2.76)} = \sqrt{4(2.76)} = \sqrt{11.04} = 3.323.$$

Ponieważ obliczona przez nas wartość statystyki $S = 11.202 > S_{0.05} = 3.323$, możemy odrzucić hipotezę zerową H_0: $\frac{\mu_2 + \mu_3 + \mu_4}{3} - \mu_5 = 0$.

Posługując się tą samą procedurą obliczeń możemy przetestować także inne równości („kontrasty"), np.:

1. H_0: $\mu_1 \frac{\mu_2 + \mu_3 + \mu_4}{3} = 0$, gdzie:

$$SE = \sqrt{9.7652\left[\frac{(1)^2}{6}+\frac{\left(\frac{1}{3}\right)^2}{6}+\frac{\left(\frac{1}{3}\right)^2}{6}+\frac{\left(\frac{1}{3}\right)^2}{6}\right]} = \sqrt{9.7652\left[\left(\frac{1}{6}\right)+\left(\frac{1}{54}\right)+\left(\frac{1}{54}\right)+\left(\frac{1}{54}\right)\right]} =$$

$$= \sqrt{9.7652\left[\frac{12}{54}\right]} = \sqrt{2.17} = 1.473$$

$$\text{oraz } S = \frac{\left|32.1-\left(\frac{40.2}{3}\right)-\left(\frac{44.1}{3}\right)-\left(\frac{41.1}{3}\right)\right|}{1.473} = \frac{|32.1-13.4-14.7-13.7|}{1.473} = \frac{|-9.7|}{1.473} = 6.585.$$

Podobnie jak poprzednio, ponieważ obliczona przez nas wartość statystyki $S = 6.585 > S_{0.05} = 3.323$, możemy odrzucić hipotezę zerową H_0: $\mu_1 - \frac{\mu_2 + \mu_3 + \mu_4}{3} = 0$.

2. H_0: $\frac{\mu_1+\mu_5}{2} - \frac{\mu_2+\mu_3+\mu_4}{3} = 0$, gdzie:

$$SE = \sqrt{9.7652\left[\frac{\left(\frac{1}{2}\right)^2}{6}+\frac{\left(\frac{1}{2}\right)^2}{6}+\frac{\left(\frac{1}{3}\right)^2}{6}+\frac{\left(\frac{1}{3}\right)^2}{6}+\frac{\left(\frac{1}{3}\right)^2}{6}\right]} = \sqrt{9.7652\left[\left(\frac{1}{24}\right)+\left(\frac{1}{24}\right)+\left(\frac{1}{54}\right)+\left(\frac{1}{54}\right)+\left(\frac{1}{54}\right)\right]} =$$

$$= \sqrt{1.3562778} = 1.1646$$

$$\text{oraz } S = \frac{\left|\left(\frac{32.1+58.3}{2}\right)-\left(\frac{40.2+44.1+41.1}{3}\right)\right|}{1.1646} = \frac{|45.2-41.8|}{1.1646} = \frac{|3.4|}{1.1646} = 2.919$$

W tym przypadku, obliczona przez nas wartość statystyki $S = 2.919 < S_{0.05} = 3.323$, zatem nie mamy podstaw aby odrzucić hipotezę zerową H_0: $\frac{\mu_1+\mu_5}{2} - \frac{\mu_2+\mu_3+\mu_4}{3} = 0$.

3. H_0: $\frac{\mu_1+\mu_4}{2} - \frac{\mu_2+\mu_3}{2} = 0$, gdzie:

$$SE = \sqrt{9.7652\left[\frac{\left(\frac{1}{2}\right)^2}{6}+\frac{\left(\frac{1}{2}\right)^2}{6}+\frac{\left(\frac{1}{2}\right)^2}{6}+\frac{\left(\frac{1}{2}\right)^2}{6}\right]} = \sqrt{9.7652\left[\left(\frac{1}{24}\right)+\left(\frac{1}{24}\right)+\left(\frac{1}{24}\right)+\left(\frac{1}{24}\right)\right]} =$$

$$= \sqrt{9.7652\left[\frac{1}{6}\right]} = \sqrt{1.6275} = 1.276$$

$$\text{oraz } S = \frac{\left|\left(\frac{32.1+41.1}{2}\right)-\left(\frac{40.2+44.1}{2}\right)\right|}{1.276} = \frac{|36.6-42.15|}{1.276} = \frac{|-5.55|}{1.276} = 4.35$$

Ponieważ $S = 4.35 > S_{0.05} = 3.323$, możemy odrzucić hipotezę zerową H_0: $\frac{\mu_1+\mu_4}{2} - \frac{\mu_2+\mu_3}{2} = 0$.

Ogólnie, w metodzie Scheffe'go średnie zestawia się zawsze na zasadzie kontrastu (równania) – dlatego o metodzie tej mówi się jako o analizie kontrastów, przy czym niejako „w opozycji do siebie" występują nie tyle same średnie, co ich bardziej złożone układy (sumy, iloczyny, ilorazy, itd.) – dlatego używamy pojęcia „kontrasty wielokrotne".

Testy badania jednorodności wariancji

Przykład 46

W tabeli podano wyniki czasu okluzji (*s*) zebrane w dwóch grupach pacjentów poddawanych terapii przeciwpłytkowej. Należy ocenić czy zmienność ocenianego parametru (czas okluzji) jest jednakowa w obu grupach pacjentów.

	grupa 1		grupa 2	
	84	79	92	96
	89	82	95	92
	89	85	90	99
	84	90	89	95
	80	83	93	93
	81	85	96	90
	86	84	94	88
	86	97	89	89
	88	81	93	91
	81	88	95	94
$\bar{x}_i$	85.100		92.650	
s_i	4.266		2.925	
s_i^2	18.20		8.555	

Hipotezy:
H_0: $\sigma_1^2 = \sigma_2^2$
H_A: $\sigma_1^2 \neq \sigma_2^2$
można zweryfikować stosując test ilorazu wariancji (znany także jako test Fishera-Snedecora). Wartość statystyki *F* tego testu liczymy jako:

$$F = \frac{s_1^2}{s_2^2} \quad \text{lub} \quad F = \frac{s_2^2}{s_1^2} \quad (F \text{ musi być większe od jedności})$$

Wartość *F* istotnie większa od 1 upoważnia nas do odrzucenia hipotezy zerowej H_0: $\sigma_1^2 = \sigma_2^2$. Wartość *F* liczona dla naszego przykładu wynosi:

$$F = \frac{18.2}{8.555} = 2.1274 \qquad F_{0.05(2),19,19} = 2.53$$

Ponieważ $F = 2.1274 < F_{0.05(2),19,19} = 2.53$, nie mamy podstaw do odrzucenia hipotezy zerowej H_0: $\sigma_1^2 = \sigma_2^2$. Skoro przyjęliśmy, że σ_1^2 oraz σ_2^2 nie są różne, a więc są reprezentatywne dla σ^2 populacji ogólnej, możemy obliczyć całkowitą (sumaryczną) wariancję obu prób:

$$s_p^2 = \frac{SS_1 + SS_2}{v_1 + v_2} = \frac{v_1 s_1^2 + v_2 s_2^2}{v_1 + v_2} = 13.3776 \qquad v_1 = n_1 - 1 = 20 - 1 = 19 \qquad v_2 = n_2 - 1 = 19$$

gdzie SS_1 i SS_2 oznaczają odpowiednie sumy kwadratów, zaś n_1 i n_2 liczebności prób.

Przykład 47

Dla danych z przykładu 42 należy ocenić jednorodność wariancji. Jednym z najczęściej stosowanych w takich przypadkach testów jest test Bartletta, którego statystyka B posiada rozkład zbliżony do rozkładu chi²:

$$B = \left(\ln s_p^2\right) * \left(\sum_{i=1}^{k} v_i\right) - \sum_{i=1}^{k} v_i \ln s_i^2 \quad \text{lub}$$

$$B = 2.30259\left[\left(\log s_p^2\right) * \left(\sum_{i=1}^{k} v_i\right) - \sum_{i=1}^{k} v_i \log s_i^2\right]$$

gdzie s_p^2 jest całkowitą wariancją obliczaną jako $\sum_{i=1}^{k} SS_i / \sum_{i=1}^{k} v_i$, n_i jest liczebnością próby i oraz $v_i = n_i - 1$.

Lepszą aproksymację rozkładu statystyki B do rozkładu chi² uzyskuje się stosując poprawkę:

$$C = 1 + \frac{1}{3(k-1)}\left(\sum_{i=1}^{k} \frac{1}{v_i} - \frac{1}{\sum_{i=1}^{k} v_i}\right) \qquad B_C = \frac{B}{C}$$

Dla $k = 2$ oraz $n_1 = n_2$ test Bartletta odpowiada testowi Fishera-Snedecora.
Testujemy parę hipotez:
H_0: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2$
H_A: wariancje w próbach nie są równe

Obliczenia:

	lek 1	**lek 2**	**lek 3**	**lek 4**	
	61	69	103	88	
	57	68	102	84	
	65	74	100	83	
	59	66	97	86	
	62	70		90	
SS_i	36.8	35.2	21.0	32.8	$\sum SS_i = 125.8$
s_i	3.03315	2.966479	2.645751	2.863564	
s_i^2	9.20	8.80	7.00	8.20	
n_i	5	5	4	5	$\sum n_i = 19$
v_i	4	4	3	4	$\sum v_i = 15$
$\log s_i^2$	0.96379	0.9445	0.8451	0.9138	
$v_i \log s_i^2$	3.85515	3.7779	2.5353	3.6553	$\sum v_i \log s_i^2 = 13.8236$
$1/v_i$	0.250	0.250	0.333	0.250	$\sum (1/v_i) = 1.0833$

$$s_p^2 = \frac{\sum SS_i}{\sum v_i} = \frac{125.8}{15} = 8.3867 \qquad \log s_p^2 = 0.92359$$

$$B = 2.30259\left[\left(\log s_p^2\right)\left(\sum v_i\right) - \sum v_i \log s_i^2\right] = 2.30259[(0.92359)(15) - 13.8236] =$$

$$= 2.30259(0.0302407) = 0.069632$$

$$C = 1 + \frac{1}{3(k-1)}\left(\sum \frac{1}{v_i} - \frac{1}{\sum v_i}\right) = 1 + \frac{1}{3(3)}\left(1.0833 - \frac{1}{15}\right) = 1.1129626$$

$$B_C = \frac{B}{C} = \frac{0.069632}{1.1129626} = 0.06228$$

Ponieważ $\chi^2_{0.05,3} = 7.815$, nie odrzucamy H_0.

Wniosek

Wariancje w próbach nie są istotnie różne.

Testy istotności do oceny biozgodności (leków)

Przykład 48

Porównywano skuteczność nowej statyny oraz dotychczas stosowanego preparatu w obniżaniu stężenia cholesterolu w osoczu krwi. Do badań wyselekcjonowano 40 pacjentów z chorobą niedokrwienną serca i badania wykonano w modelu krzyżowym. W drugiej fazie doświadczenia u jednego z pacjentów wystąpiło pogorszenie ogólnego stanu zdrowia i decyzją lekarzy został on wyłączony z dalszych badań.

Należy obliczyć przedział ufności dla testu jednostronnego do oceny biozgodności obu preparatów, przy założeniu, że różnica między skutecznością działania obu leków nie powinna przekraczać 10%.

Uzyskano następujące wartości efektów hipolipemizujących dla obu preparatów:

	stary lek (referencyjny)	**nowy lek (testowany)**
średnia	45.6	41.2
SD	3.7	4.1
n	40.0	39.0

10% różnica wynosi: δ = 45.6x0.10 = 4.56, natomiast obserwowana różnica: d = 41.2 – 45.6 = – 4.4

Obliczamy wartość błędu standardowego:

$$s_p^2 = \frac{s_1^2(n_1-1)+s_2^2(n_2-1)}{n_1+n_2-2)} =$$

$$= \frac{(40-1)(3.7)^2+(39-1)(4.1)^2}{40+39-2} = \frac{39*13.69+38*16.81}{77} = 15.23$$

$$SE = \sqrt{\frac{s_p^2}{n_1}+\frac{s_p^2}{n_2}} = \sqrt{\frac{15.23}{40}+\frac{15.23}{39}} = 0.8782$$

$$\text{kres dolny} = \frac{(d-SE)+\bar{x}_{referencyjne}}{\bar{x}_{referencyjne}} \times 100\% = \frac{(-4.4-0.8782)+45.6}{45.6} * 100 = 88.42\%$$

$$\text{kres górny} = \frac{(d+SE)+\bar{x}_{referencyjne}}{\bar{x}_{referencyjne}} \times 100\% = \frac{(-4.4+0.8782)+45.6}{45.6} * 100 = 92.3\%$$

Ostatecznie nasz przedział ufności będzie wynosił:

$$88.4\% < \frac{\mu_{testowane}}{\mu_{referencyjne}} < 92.3\%$$

Rzeczywista różnica między dwoma preparatami leży w przedziale 88.4 – 92.3%, a zatem wykracza poza dopuszczalne 10% kryterium różnicy między dwoma lekami. Na podstawie oszacowanego przedziału ufności nie możemy zatem uznać, że dwa porównywane preparaty są biozgodne.

Stosując dwa testy jednostronne sprawdzamy, czy różnica między dwoma preparatami przekracza 10%.
Nasze hipotezy mają postać:

H_{01}: $\mu_{nowy} - \mu_{stary} \leq -10\%$
H_{A1}: $\mu_{nowy} - \mu_{stary} > -10\%$

oraz
H_{02}: $\mu_{nowy} - \mu_{stary} \geq +10\%$
H_{A2}: $\mu_{nowy} - \mu_{stary} < +10\%$

$\delta = 45.6 \text{x} 0.10 = 4.56$
$d = 41.2 - 45.6 = -4.4$

Dla poziomu istotności $\alpha = 0.05$ odrzucimy hipotezy zerowe H_{01} oraz H_{02} jeżeli

$$|t_{dośw}| > t_{0.05(2),77} = 1.665 \qquad t_1 = \frac{\left(\bar{x}_{nowy} - \bar{x}_{stary}\right) - \delta}{SE} = \frac{-4.4 - (-4.56)}{0.8782} = 0.1822$$

Ponieważ $|t_{dośw}| > t_{0.05(2),77} = 1.665$, nie mamy podstaw aby odrzucić H_{01}: $\mu_{nowy} - \mu_{stary} \leq -10\%$.

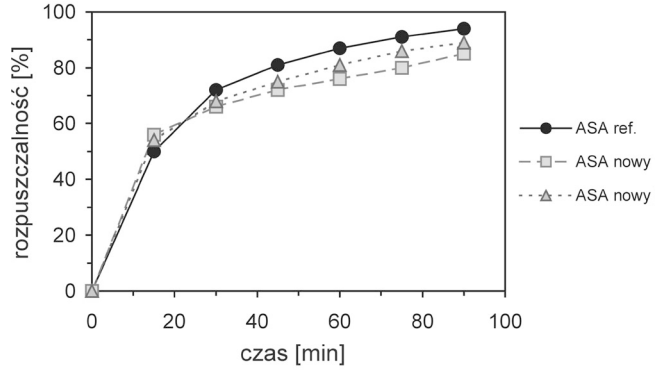
$$t_2 = \frac{\delta - \left(\bar{x}_{nowy} - \bar{x}_{stary}\right)}{SE} = \frac{4.56 - (-4.4)}{0.8782} = 10.203$$

Ponieważ $|t_{dośw}| > t_{0.05(2),77} = 1.665$, możemy odrzucić H_{02} mówiącą, że różnica między preparatami nie jest mniejsza niż 10%, i przyjąć H_{A2}: $\mu_{nowy} - \mu_{stary} < +10\%$.

Możemy zatem wnioskować z prawdopodobieństwem 95%, że aktywność nowego leku zawiera się w przedziale 90-110% w stosunku do leku referencyjnego.

Przykład 49

Badano profile rozpuszczalności dwóch nowych preparatów kwasu acetylosalicylowego i porównywano je z preparatem dostępnym na rynku. Profile czasowe rozpuszczalności przedstawiono na wykresie, a średnie wartości frakcji rozpuszczonych podano w tabeli.



czas	ASA ref.	ASA nowy 1	ASA nowy 2
0	0	0	0
15	50	56	54
30	72	66	68
45	81	72	75
60	87	76	81
75	91	80	86
90	94	85	89

Jakie jest podobieństwo każdego z nowych preparatów w stosunku do leku referencyjnego?

Obliczenia:

czas	R_t	T_{t1}	T_{t2}	R_t-T_{t1}	$(R_t-T_{t1})^2$	R_t-T_{t2}	$(R_t-T_{t2})^2$
0	0	0	0				
15	50	56	54	–6	36	–4	16
30	72	66	68	6	36	4	16
45	81	72	75	9	81	6	36
60	87	76	81	11	121	6	36
75	91	80	86	11	121	5	25
90	94	85	89	9	81	5	25
$\sum$	475			40	476	22	154

Obliczamy współczynnik różnicy i współczynnik podobieństwa dla nowego leku 1:

$$f_{różnic} = \frac{\sum |R_t - T_t|}{\sum R_t} \times 100\% = \frac{40}{475} \times 100\% = 8.42\%$$

$$f_{podobieństw} = 50 * \log\left[\frac{1}{\sqrt{1 + \frac{1}{n}\sum (R_t - T_t)^2}} \times 100\%\right] =$$

$$= 50 * \log\left[\frac{1}{\sqrt{1 + \frac{1}{6}(476)}} \times 100\%\right] = 50 * \log[11.16] = 50 * 1.0476 = 52.38\%$$

Ponieważ $f_{różnicy} < 15\%$ i $f_{podobieństw} > 50\%$, możemy uznać, że profile rozpuszczalności leku nowego 1 i leku referencyjnego nie są istotnie różne.

Obliczamy współczynnik różnicy i współczynnik podobieństwa dla nowego leku 2:

$$f_{różnic} = \frac{\sum |R_t - T_t|}{\sum R_t} \times 100\% = \frac{22}{475} \times 100\% = 4.63\%$$

$$f_{podobieństw} = 50 * \log\left[\frac{1}{\sqrt{1 + \frac{1}{n}\sum (R_t - T_t)^2}} \times 100\%\right] =$$

$$= 50 * \log\left[\frac{1}{\sqrt{1+\frac{1}{6}(154)}} \times 100\%\right] = 50 * \log[19.365] = 50 * 1.287 = 64.35\%$$

Ponieważ $f_{różnicy} < 15\%$ i $f_{podobieństw} > 50\%$, możemy uznać, że profile rozpuszczalności leku nowego 2 i leku referencyjnego nie są istotnie różne.

Przykład 50

W celu analizy składu nowego preparatu krwiotwórczego próbkę leku rozdzielono na dwie części oraz wysłano do dwóch niezależnych pracowni analitycznych. Każda z pracowni przeprowadziła po 50 niezależnych analiz próbki i uzyskała następujące wyniki zawartości aktywnego składnika:

	pracownia analityczna 1	**pracownia analityczna 2**
średnia	94.67%	96.41%
wariancja	25.71	22.79
SD	5.07%	4.77%
n	50	50

1. Czy różnica między pracowniami jest istotna?
2. Czy można uznać, że różnica między pracowniami wynosi mniej niż 5%?
3. Czy liczebność próbek jest wystarczająco duża, aby wykryć 5% różnicę z mocą wnioskowania równą 80%?
4. Zakładając, że oczekujemy różnicy między pracowniami mniejszej niż 5%, jaki będzie przedział ufności określający biozgodność?

1. Obliczamy czy różnica między pracowniami jest istotna.

$$s_p^2 = \frac{s_1^2(n_1-1)+s_2^2(n_2-1)}{(n_1+n_2-2)} =$$

$$\frac{(50-1)(25.71)+(50-1)(22.79)}{50+50-2} = \frac{49*25.71+49*22.79}{98} = 24.25$$

$$\text{i odpowiednio, } SE = \sqrt{\frac{2s_p^2}{n}} = \sqrt{\frac{2*24.25}{50}} = 0.985$$

Para hipotez do obustronnego testu *t* ma postać:
H_0: $\mu_1 = \mu_2$
H_A: $\mu_1 \neq \mu_2$

Zakładając poziom istotności $\alpha = 0.05$, będziemy mogli odrzucić H_0: $\mu_1 = \mu_2$, jeżeli $|t_{dośw}| > t_{0.05(2),98} = 1.984$.

$$t = \frac{\bar{x}_1 - \bar{x}_2}{SE} = \frac{94.67 - 96.41}{0.985} = \frac{-1.74}{0.985} = -1.7666$$

Ponieważ $|t_{dośw}| < t_{0.05(2),98} = 1.984$, nie ma podstaw do odrzucenia H_0 mówiącej, że nie ma różnicy między pracowniami. Wnioskujemy zatem, że różnica między pracowniami nie jest istotna.

2. Stosując dwa testy jednostronne obliczamy czy możemy uznać, że różnica między pracowniami wynosi mniej niż 5%.

Nasze pary hipotez będą miały postać:

$H_{01}: \mu_{pracownia\ 1} - \mu_{pracownia\ 2} \leq -5\%$
$H_{A1}: \mu_{pracownia\ 1} - \mu_{pracownia\ 2} > -5\%$

oraz

$H_{02}: \mu_{pracownia\ 1} - \mu_{pracownia\ 2} \geq +5\%$
$H_{A2}: \mu_{pracownia\ 1} - \mu_{pracownia\ 2} < +5\%$

$\mu_{pracownia\ 1} - \mu_{pracownia\ 2} = -1.74\%$

$\delta = 0.05 \text{ x } 94.67 = 4.7335$

Dla poziomu istotności $\alpha = 0.05$ odrzucimy nasze hipotezy zerowe H_{01} oraz H_{02} jeżeli $|t_{dośw}| > t_{0.05(2),98} = 1.984$

$$t_1 = \frac{\left(\bar{x}_{nowy} - \bar{x}_{stary}\right) - \delta}{SE} = \frac{(94.67 - 96.41) - (-94.67 * 0.05)}{0.985} = 3.04$$

Ponieważ $|t_{dośw}| > t_{0.05(2),98} = 1.984$, możemy odrzucić H_{01} mówiącą, że różnica między preparatami nie jest większa niż –5%.

$$t_2 = \frac{\delta - (\bar{x}_{nowy} - \bar{x}_{stary})}{SE} = \frac{4.735 - (-1.74)}{0.985} = 6.572$$

Ponieważ $|t_{dośw}| > t_{0.05(2),98} = 1.984$, możemy odrzucić H_{01} mówiącą, że różnica między preparatami nie jest mniejsza niż +5%.

Zatem z prawdopodobieństwem 95% możemy wnioskować, że różnica między pomiarami wykonanymi w obu pracowniach zawiera się w przedziale 95-105%.

3. Obliczamy jaka powinna być liczebność próbek, aby wykryć 5% różnicę (jeżeli taka istnieje) z mocą wnioskowania równą 80%?

Przy założeniu, że pracownia 1 jest pracownią referencyjną, 5% różnica między wynikami uzyskanymi w obu pracowniach wynosi $d = 94.67 \text{ x } 0.05 = 4.7335$.

Skoro wartość krytyczna statystyki testu t Studenta dla poziomu istotności $\alpha = 0.05$ wynosi $t_{0.05(2),49} = 2.010$, to:

$$t_{\beta(1),n-1} \geq \frac{\delta}{SE} - t_{\alpha,n-1} \qquad t_\beta \geq \frac{4.7335}{0.985} - 2.01 = 2.796$$

Takiej wartości statystyki t_β dla *df* = 49 stopni swobody odpowiada istotność β = 0.01, stąd możemy wnioskować, że moc testu wynosi 1 – 0.01 = 0.99.

Odwróćmy teraz problem i zapytajmy: ile powtórzeń potrzebowalibyśmy wykonać, aby z mocą przynajmniej 80% wykryć różnicę między pracowniami wynoszącą 5%?

Zapiszmy powyższe równanie w innej postaci i przekształćmy je:

wiedząc, że $SE = \sqrt{\frac{2s_p^2}{n}}$ możemy zapisać: $\delta = \sqrt{\frac{s^2}{n}}(t_{\alpha,n-1} + t_{\beta(1),n-1})$

Stąd: $n \geq \frac{2s_p^2}{\delta^2}(t_{\alpha,n-1} + t_{\beta(1),n-1})$

Skoro wiemy, że:

$s_p^2 = 24.25$

$t_{0.05,49} = 2.010$

$t_{0.20(1),49} = 0.85$

$\delta = 4.7335$

mamy:

$$n \geq \frac{2(24.25)}{(4.7335)^2}(0.85 + 2.01) = \frac{48.5}{22.406}(2.86) = 6.2$$

Zaledwie po 7 pomiarów z każdej pracowni wystarczy, aby z mocą testu 80% wykryć 5% różnicę wyników analizy.

4. Skoro oczekujemy różnicy między pracowniami mniejszej niż 5%, obliczamy jaki będzie przedział ufności określający biozgodność.

$$\text{kres dolny} = \frac{(d - SE) + \bar{x}_{referencyjne}}{\bar{x}_{referencyjne}} \times 100\% = \frac{(-1.73 - 0.985) + 94.67}{94.67} * 100 = 97.13\%$$

$$\text{kres górny} = \frac{(d + SE) + \bar{x}_{referencyjne}}{\bar{x}_{referencyjne}} \times 100\% = \frac{(-1.73 + 0.985) + 94.67}{94.67} * 100 = 99.2\%$$

Ostatecznie nasz przedział ufności będzie wynosił:

$$97.13\% < \frac{\mu_{testowane}}{\mu_{referencyjne}} < 99.2\%$$

Granice przedziału mieszczą się w granicach przyjętego kryterium (100 ± 5%), zatem możemy wnioskować, że wyniki uzyskiwane w obu pracowniach są wystarczająco zgodne.

Przykład 51

W tabeli przedstawiono stopień hamowania agregacji płytek krwi dla różnych stężeń blokera receptora dla fibrynogenu w dwóch badanych grupach: zdrowych dawców oraz pacjentów. Należy określić czy występuje podobieństwo we wrażliwości płytek krwi na działanie blokera u pacjentów i zdrowych dawców.

Zdrowi dawcy krwi

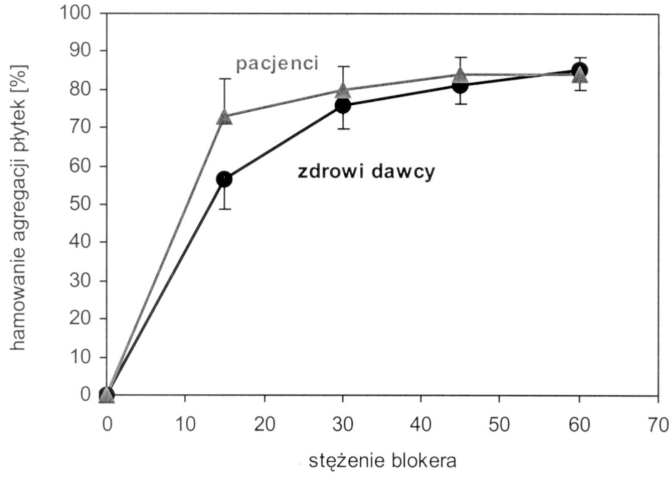
stężenie	**hamowanie agregacji (%)**						**średnia**	**SD**	**CV**
15	64	43	58	62	53	59	56.50	7.61	13.47
30	84	76	77	79	75	65	76.00	6.26	8.24
45	85	75	83	76	80	88	81.17	5.12	6.30
60	91	87	86	75	84	89	85.33	5.61	6.57

Pacjenci

stężenie	**hamowanie agregacji (%)**						**średnia**	**SD**	**CV**
15	79	67	57	79	72	84	73.00	9.86	13.51
30	73	84	85	80	72	86	80.00	6.16	7.71
45	79	85	90	83	80	88	84.17	4.36	5.17
60	78	85	90	83	80	88	84.00	4.60	5.48

Obliczamy sumy wartości średnich dla poszczególnych stężeń blokera:

stężenie	**zdrowi (Z)**	**pacjenci (P)**	**(Z-P)**	**$(Z-P)^2$**
15	56.50	73.00	16.50	272.25
30	76.00	80.00	4.00	16.00
45	81.17	84.17	3.00	9.00
60	85.33	84.00	1.33	1.78
$\sum$	299.00	321.17	24.83	299.03



Obliczamy współczynnik różnicy i współczynnik podobieństwa:

$$f_{różnic} = \frac{\sum |R_t - T_t|}{\sum R_t} \times 100\% = \frac{24.83}{299.0} \times 100\% = 8.3\%$$

$$f_{podobieństw} = 50 * \log\left[\frac{1}{\sqrt{1 + \frac{1}{n}\sum (R_t - T_t)^2}} \times 100\%\right] =$$

$$= 50 * \log\left[\frac{1}{\sqrt{1 + \frac{1}{4}(299.03)}} \times 100\%\right] = 50 * \log[11.49] = 50 * 1.0603 = 53.01\%$$

Ponieważ $f_{różnicy} < 15\%$ i $f_{podobieństw} > 50\%$, możemy wnioskować, że krzywe hamowania agregacji płytek przez bloker w dwóch grupach pacjentów nie są istotnie różne.

Ocena zgodności dwóch metod

Przykład 52

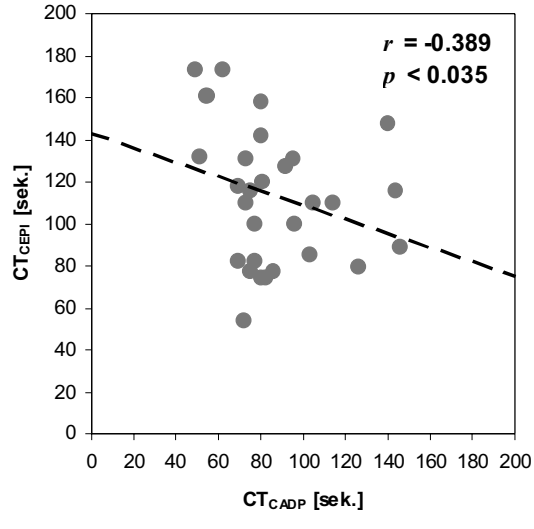
W analizatorze przepływowym PFA-100™, który pozwala badać reaktywność płytek krwi w warunkach zbliżonych do naturalnych, zastosowano dwa rodzaje kaset pomiarowych z układem przepływowym symulującym uszkodzone naczynie krwionośne. Obie kasety zawierają kapilarny układ przepływowy, a różnią się rodzajem membrany aktywującej: w jednej wersji zawiera ona kolagen i epinefrynę (CEPI), w drugiej – kolagen i ADP (CADP). Płytki krwi przepływając przez naczynie aktywują się na powierzchni membrany i powodują zamknięcie jej otworu przepływowego, odnotowane przez analizator jako czas okluzji (CT). W tabeli poniżej przedstawiono wyniki czasu okluzji zmierzonego dwukrotnie przy użyciu kaset z kolagenem i epinefryną (CT_{CEPI}) oraz kaset z kolagenem i ADP (CT_{CADP}). Należy ocenić czy pomiary czasu okluzji uzyskiwane przy użyciu obu typów kaset są zgodne oraz jaka jest powtarzalność pomiarów dla każdego rodzaju kaset.

dawca	CT_{CADP} (sek.)		CT_{CEPI} (sek.)		różnica (CT_{CADP}–CT_{CEPI}) (sek.)		średnia (CT_{CADP}; CT_{CEPI}) (sek.)	
	pomiar pierwszy	pomiar drugi	pomiar pierwszy	pomiar drugi	pomiar pierwszy	pomiar drugi	pomiar pierwszy	pomiar drugi
1	140.0	151.0	148.0	140.0	–8.0	11.0	144.0	145.5
2	126.0	106.0	79.0	88.0	47.0	18.0	102.5	97.0
3	144.0	134.0	116.0	104.0	28.0	30.0	130.0	119.0
4	69.0	69.0	118.0	107.0	–49.0	–38.0	93.5	88.0
5	69.0	61.0	82.0	85.0	–13.0	–24.0	75.5	73.0

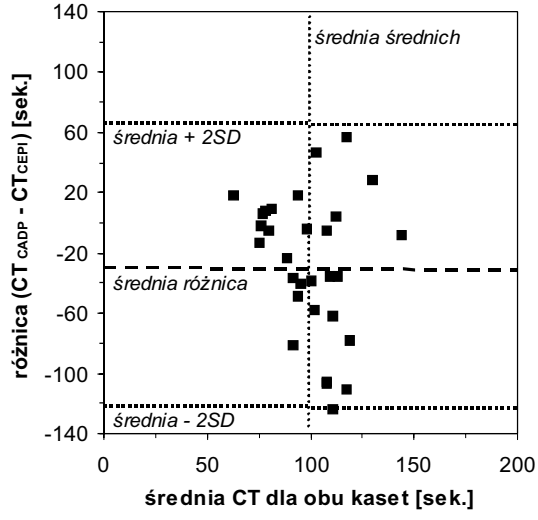
6	81.0	86.0	120.0	114.0	–39.0	–28.0	100.5	100.0
7	105.0	83.0	110.0	118.0	–5.0	–35.0	107.5	100.5
8	103.0	96.0	85.0	79.0	18.0	17.0	94.0	87.5
9	114.0	143.0	110.0	118.0	4.0	25.0	112.0	130.5
10	77.0	70.0	100.0	115.0	–23.0	–45.0	88.5	92.5
11	80.0	81.0	158.0	182.0	–78.0	–101.0	119.0	131.5
12	80.0	72.0	142.0	181.0	–62.0	–109.0	111.0	126.5
13	49.0	68.0	173.0	182.0	–124.0	–114.0	111.0	125.0
14	55.0	77.0	161.0	165.0	–106.0	–88.0	108.0	121.0
15	80.0	80.0	74.0	65.0	6.0	15.0	77.0	72.5
16	72.0	68.0	54.0	112.0	18.0	–44.0	63.0	90.0
17	86.0	77.0	77.0	86.0	9.0	–9.0	81.5	81.5
18	73.0	66.0	131.0	141.0	–58.0	–75.0	102.0	103.5
19	146	106.0	89.0	96.0	57.0	10.0	117.5	101.0
20	75.0	82.0	116.0	132.0	–41.0	–50.0	95.5	107.0
21	51.0	49.0	132.0	158.0	–81.0	–109.0	91.5	103.5
22	96.0	72.0	100.0	109.0	–4.0	–37.0	98.0	90.5
23	92.0	59.0	127.0	145.0	–35.0	–86.0	109.5	102.0
24	62.0	51.0	173.0	182.0	–111.0	–131.0	117.5	116.5
25	54.0	48.0	161.0	165.0	–107.0	–117.0	107.5	106.5
26	82.0	120.0	74.0	65.0	8.0	55.0	78.0	92.5
27	73.0	64.0	110.0	54.0	–37.0	10.0	91.5	59.0
28	75.0	84.0	77.0	86.0	–2.0	–2.0	76.0	85.0
29	95.0	125.0	131.0	141.0	–36.0	–16.0	113.0	133.0
30	77.0	62.0	82.0	85.0	–5.0	–23.0	79.5	73.5
$\bar{x}$	86.0	83.7	113.7	120.0	**–27.6**	–36.3	**99.9**	101.8
s	26.2	27.5	32.7	38.0	**47.2**	51.7	**17.9**	20.8

Do oceny zgodności wyników wykorzystamy dane umieszczone w kolumnach „pierwszy pomiar", zaś dane z kolumn „drugi pomiar" posłużą nam do analizy powtarzalności.

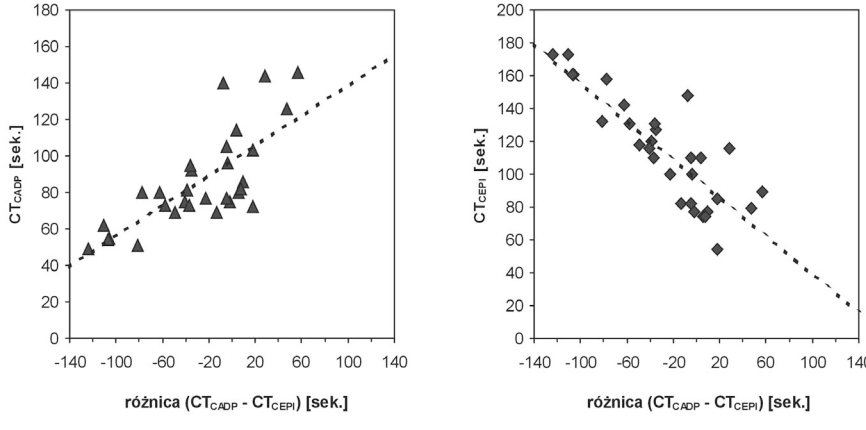
Prosty wykres zależności wyników uzyskanych dla każdej z kaset pomiarowych nie wskazuje nam w jasny sposób do jakiego stopnia wyniki te są zgodne czy niezgodne. Jak wskazuje wartość współczynnika prostej korelacji liniowej, obie zmienne (CT_{CEPI} i CT_{CADP}) są w istotny sposób zależne jedna od drugiej, ale trudno nam się wypowiedzieć, czy zależność taka jest wystarczającym czy niedostatecznym dowodem zgodności danych.



Miarę zgodności, czy niezgodności pomiarów o wiele lepiej odzwierciedla wykres zależności różnic (d) między wynikami dla obu kaset (różnica [$CT_{CADP} - CT_{CEPI}$]) względem średnich ($\bar{x}$) wyników dla obu kaset (średnia [CT_{CADP}, CT_{CEPI}]).



Różnice wartości wyników uzyskanych przy użyciu kaset CEPI i kaset CADP sięgają 95 sek., co wskazuje, iż wyniki uzyskane przy użyciu dwóch rodzajów kaset nie są bardzo zgodne. To co w oczywisty sposób wynika z tego wykresu, nie było tak łatwo dostrzegalne na pierwszym wykresie zależności CT_{CEPI} i CT_{CADP}. Zauważmy, że powyższy wykres jest niczym innym, jak zależnością między błędem pomiarowym (w naszym przypadku są to różnice między wynikami dla kasety CEPI i wynikami dla kasety CADP) oraz oszacowaną wartością rzeczywistą (w naszym przypadku średnia ze średnich ($\bar{x}$) dla wyników uzyskanych przy zastosowaniu obu rodzajów kaset jest najlepszą z możliwych szacowaną wartością rzeczywistą próby). Dla porównania, spójrzmy także na zależności różnic wyników zebranych dla obu kaset oraz poszczególnych wartości dla każdej z kaset.



Proporcjonalne zależności obserwowane w takim przypadku są oczywiste, ponieważ różnica między zmiennymi będzie zawsze silnie zależała od każdej pojedynczej zmiennej, ale zależność ta nie mówi nam nic na temat zgodności między tymi zmiennymi.

Dla naszego porównania średnia różnica (d) między dwoma kasetami wynosi $\overline{d} = -27.6$ sek., zaś odchylenie standardowe różnic $s = 47.2$ sek. Znając średnią różnicę wyników uzyskiwanych dla dwóch porównywanych typów kaset możemy pokusić się o aproksymację (to znaczy wzajemne przeliczanie) wyników uzyskanych jedną metodą (z wykorzystaniem jednej kasety) na wyniki uzyskiwane metodą drugą (z użyciem drugiej kasety). W naszym przypadku możemy powiedzieć, że wynik uzyskiwany przy użyciu kaset CADP będzie średnio o 27.6 sek. niższy od wyniku uzyskanego przy wykorzystaniu kaset CEPI. Należy jednak zauważyć, że aproksymacja taka będzie dobra jedynie wtedy gdy rozrzut tych różnic będzie niewielki, mieszczący się w niewielkim przedziale ufności wokół średniej różnic. Dla naszego porównania wartość odchylenia różnic jest wysoka (47.6 sek.) i wskazuje, że większość obserwowanych różnic między wynikami uzyskiwanymi przy użyciu dwóch kaset będzie leżała w szerokim przedziale wartości od $\overline{d} - 2 * s = -27.6 - 2*(47.2) = -122.1$ sek. do $\overline{d} + 2 * s = -27.6 + 2*(47.2) = 66.9$ sek. Wynika stąd, że wartości CT mierzone przy użyciu kaset CEPI mogą być o 140% wyższe lub o prawie 80% niższe od wyników zbieranych przy użyciu kaset CADP, dla których średnia wartość parametru CT_{CADP} w badanej próbie wynosi 86 sek.

Zdajemy sobie sprawę, że oszacowana przez nas miara niezgodności między porównywanymi metodami odnosi się jedynie do małej przebadanej próby losowej, ale niekoniecznie musi być reprezentatywna dla populacji ogólnej. Toteż powinniśmy jeszcze obliczyć błąd, z jakim zostały oszacowane średnia różnica oraz granice przedziału jej zmienności.

Skoro odchylenie standardowe różnic wynosi $s = 47.2$ sek., to wartość błędu standardowego średniej różnicy ($\overline{d}$) między kasetami możemy obliczyć jako $SE_{\overline{d}} = \sqrt{(s^2 / n)} = 8.63$, zaś błąd standardowy granicy dolnej lub górnej jako $SE_{\overline{d}-2s,\overline{d}+2s} = \sqrt{(3s^2 / n)} = 14.94$. Dla 95% przedziału ufności ($95\%CI$) oraz $n - 1 = 30 - 1 = 29$ stopni swobody wartość krytyczna obustronna rozkładu t Studenta wynosi $t_{0.05,29} = 2.04$. Zatem, 95% przedział ufności dla niezgodności (różnic) obu metod wynosi:

od –27.6 – (2.04 x 8.63) = –45.2 sek. do –27.6 + (2.04 x 8.63) = –10.0 sek.,

czyli możemy zapisać:

$95\%CI_{\overline{d}} = -27.6 \pm 17.6$ (–45.2; 10.0) sek.

95% przedziały ufności dla dolnej i górnej granicy przedziału wynoszą odpowiednio:

od –122.1 – (2.04 x 14.94) = –152.6 sek. do –122.1 + (2.04 x 14.94) = –91.6 sek.,

gdyż $95\%CI_{\overline{d}-2s} = -122.1 \pm 30.5$ (-152.6; -91.6) sek.

oraz od 66.9 – (2.04 x 14.94) = 36.4 sek. do 66.9 + (2.04 x 14.94) = 97.4 sek.,

gdyż $95\%CI_{\overline{d}+2s} = 66.9 \pm 30.5$ (36.4; 97.4) sek.

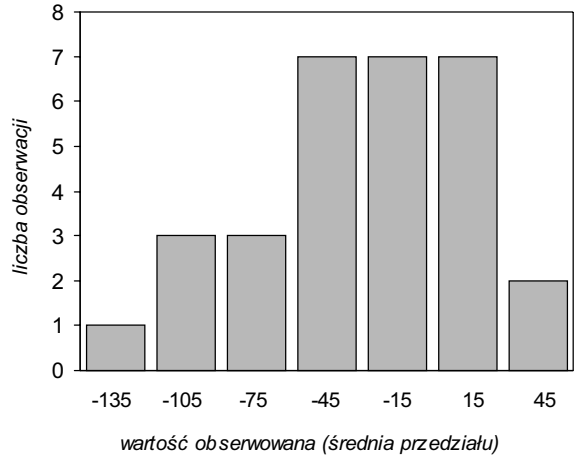
Oczywiście oszacowane przez nas granice $\overline{d} - 2 * s = -122.1$ sek. oraz $\overline{d} + 2 * s = 66.9$ sek. wyznaczają poprawnie zakres zmienności pod warunkiem, że rozkład różnic jest rozkła-

dem normalnym, czyli wtedy gdy 95% wszystkich możliwych do stwierdzenia różnic znajdzie się w zakresie:

$\overline{d} - 1.96 * s = -27.6 - 1.96*(47.2) = -120.2$ sek.

oraz $\overline{d} + 1.96 * s = -27.6 + 1.96*(47.2) = 65.0$ sek.

Czy tak jest naprawdę najłatwiej się przekonać spoglądając na histogram przedstawiający rozkład częstości różnic.



Widzimy, że rozkład różnic nie jest normalny, wykazuje wyraźną asymetrię – jest lewoskośny. W takim przypadku szacowany przedział ufności nie jest oczywiście wiarygodną miarą odchyleń od średniej różnicy ($\overline{d}$) między wynikami uzyskiwanymi przy użyciu dwóch rodzajów kaset.

Nasze obliczenia wskazują zatem, że wyniki pomiarów z wykorzystaniem kaset CEPI mogą być o około 122 sek. niższe lub o około 67 sek. wyższe od wyników zbieranych z wykorzystaniem kaset CADP. Zatem, nie jest uzasadnione twierdzić, że oba typy kaset mogą być stosowane wymiennie, gdyż kasety CADP dają z reguły o wiele niższe odczyty niż kasety CEPI.

Gdybyśmy chcieli ocenić zgodność obu metod wykorzystując wszystkie przeprowadzone powtórzenia, postępujemy nieco inaczej. Ponieważ w tym przypadku uwzględniamy więcej niż jedno powtórzenie pomiaru każdego obiektu, liczone w powyższy sposób odchylenie standardowe będzie zaniżone. Jest to zrozumiałe, gdyż procedura powtarzania pomiarów przyczynia się do zmniejszania wartości błędu systematycznego popełnianego przy prowadzeniu obserwacji. Dlatego też w przypadku takim należałoby skorygować szacowaną wartość odchylenia standardowego. Możemy to uczynić w następujący sposób. Liczymy wartości odchyleń standardowych różnic dla powtórzeń pomiarów wykonywanych dla danego obiektu badań dla każdej porównywanej metody oddzielnie. Odchylenia te, $s_{metoda\ 1}$ oraz $s_{metoda\ 2}$ odzwierciedlają zmienność wynikającą z popełnianego błędu systematycznego przy dokonywaniu pomiarów, nie zaś z różnic samych porównywanych metod. Oprócz tego obliczamy odchylenie różnic między średnimi (z powtórzeń) obliczonymi dla

każdej z metod, $s_{D_{(\bar{x}_1-\bar{x}_2)}}$. Ta wartość odchylenia odzwierciedla zarówno zmienność wynikającą z różnic między metodami, jak i zmienność w obrębie powtórzeń (spowodowaną błędem systematycznym).

Dla analizowanego przypadku mamy:

dawca	CT_{CADP} (sek.)		CT_{CEPI} (sek.)		różnica (pomiar I – pomiar II) (sek.)		średnia (pomiar I – pomiar II) (sek.)		różnica średnich (CT_{CADP} – CT_{CEPI})(sek.)
	pomiar pierwszy	pomiar drugi	pomiar pierwszy	pomiar drugi	CT_{CADP}	CT_{CEPI}	CT_{CADP} (sek.)	CT_{CEPI} (sek.)	
1	140.0	151.0	148.0	140.0	–11.0	8.0	145.5	144.0	1.5
2	126.0	106.0	79.0	88.0	20.0	–9.0	116.0	83.5	32.5
3	144.0	134.0	116.0	104.0	10.0	12.0	139.0	110.0	29.0
4	69.0	69.0	118.0	107.0	0.0	11.0	69.0	112.5	–43.5
5	69.0	61.0	82.0	85.0	8.0	–3.0	65.0	83.5	–18.5
6	81.0	86.0	120.0	114.0	–5.0	6.0	83.5	117.0	–33.5
7	105.0	83.0	110.0	118.0	22.0	–8.0	94.0	114.0	–20.0
8	103.0	96.0	85.0	79.0	7.0	6.0	99.5	82.0	17.5
9	114.0	143.0	110.0	118.0	–29.0	–8.0	128.5	114.0	14.5
10	77.0	70.0	100.0	115.0	7.0	–15.0	73.5	107.5	–34.0
11	80.0	81.0	158.0	182.0	–1.0	–24.0	80.5	170.0	–89.5
12	80.0	72.0	142.0	181.0	8.0	–39.0	76.0	161.5	–85.5
13	49.0	68.0	173.0	182.0	–19.0	–9.0	58.5	177.5	–119.0
14	55.0	77.0	161.0	165.0	–22.0	–4.0	66.0	163.0	–97.0
15	80.0	80.0	74.0	65.0	0.0	9.0	80.0	69.5	10.5
16	72.0	68.0	54.0	112.0	4.0	–58.0	70.0	83.0	–13.0
17	86.0	77.0	77.0	6.0	9.0	–9.0	81.5	81.5	0.0
18	73.0	66.0	131.0	141.0	7.0	–10.0	69.5	136.0	–66.5
19	146.0	106.0	89.0	96.0	40.0	–7.0	126.0	92.5	33.5
20	75.0	82.0	116.0	132.0	–7.0	–16.0	78.5	124.0	–45.5
21	51.0	49.0	132.0	158.0	2.0	–26.0	50.0	145.0	–95.0
22	96.0	72.0	100.0	109.0	24.0	–9.0	84.0	104.5	–20.5
23	92.0	59.0	127.0	145.0	33.0	–18.0	75.5	136.0	–60.5
24	62.0	51.0	173.0	182.0	11.0	–9.0	56.5	177.5	–121.0
25	54.0	48.0	161.0	165.0	6.0	–4.0	51.0	163.0	–112.0
26	82.0	120.0	74.0	65.0	–38.0	9.0	101.0	69.5	31.5
27	73.0	64.0	110.0	54.0	9.0	56.0	68.5	82.0	–13.5
28	75.0	84.0	77.0	86.0	–9.0	–9.0	79.5	81.5	–2.0
29	95.0	125.0	131.0	141.0	–30.0	–10.0	110.0	136.0	–26.0
30	77	62	82	85	15	–3	69.5	83.5	–14.0
średnia	86.0	83.7	113.7	120.0	2.4	–6.3	84.9	116.8	–32.0
odch. std.	26.2	27.5	32.7	38.0	17.8	18.9	25.3	34.2	47.6

Stąd $s_{metoda\,1}$ = 17.8 sek., $s_{metoda\,2}$ = 18.9 sek. oraz $s_{D_{(\bar{x}_1-\bar{x}_2)}}$ = 47.6 sek.

Skorygowaną wartość odchylenia obliczamy jako:

$$= s_{cor} = \sqrt{[s_{D_{(\bar{x}_1-\bar{x}_2)}}]^2 + \tfrac{1}{4}[s_{metoda\,1}]^2 + \tfrac{1}{4}[s_{metoda\,2}]^2}$$

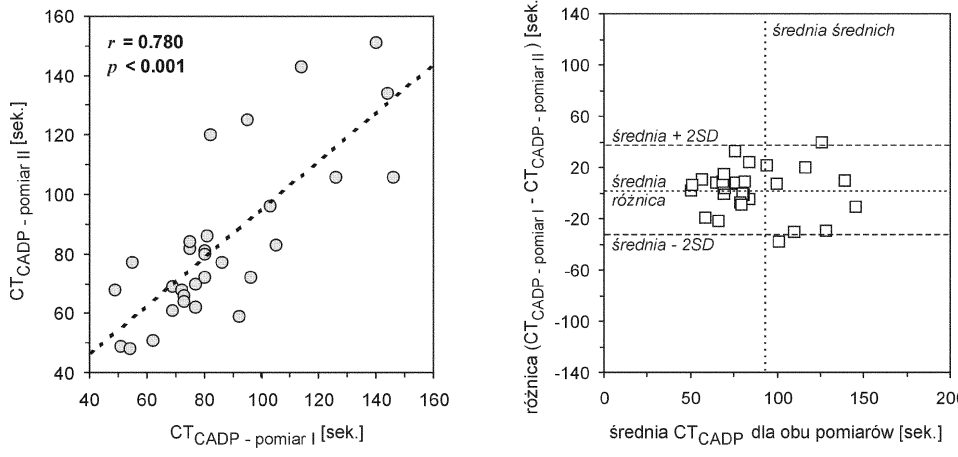
$$= \sqrt{(47.6)^2 + \tfrac{1}{4}(17.8)^2 + \tfrac{1}{4}(18.9)^2} = 49.4 \text{ sek.}$$

Wartość ta jest bardzo bliska tej, którą uzyskaliśmy szacując zgodność na podstawie pierwszego powtórzenia dla każdej metody (odchylenie standardowe różnic pierwszych pomiarów obu metod 47.2 sek.), jak również przy ocenie na podstawie drugich pomiarów dla obu metod (51.7 sek.).

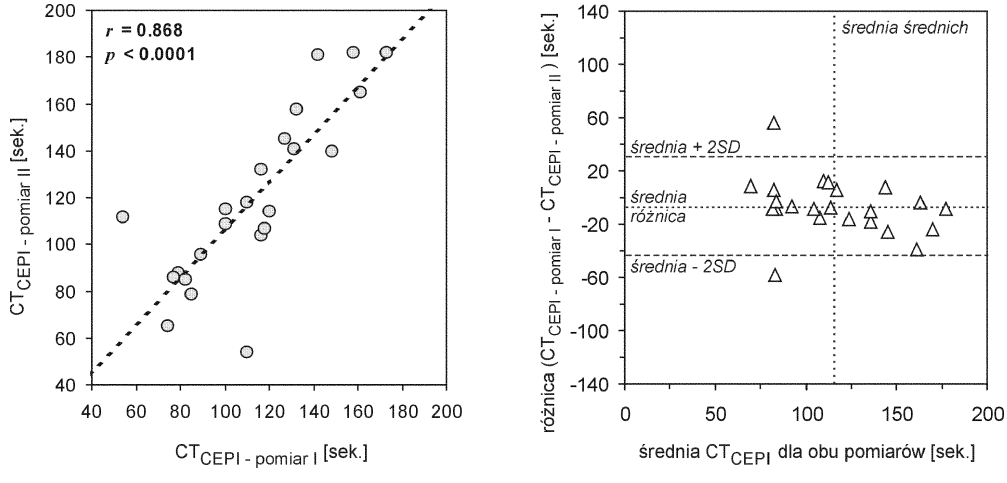
W celu oszacowania powtarzalności pomiarów przy użyciu każdej z metod wykorzystamy wyniki zebrane w sparowanych kolumnach opisanych jako *„pomiar pierwszy"* i *„pomiar drugi"*.

W sposób analogiczny do tego jaki zastosowaliśmy przy badaniu zgodności zestawmy graficzne zależności różnic (*d*) między pomiarem pierwszym i pomiarem drugim dla każdej z kaset (różnica [$CT_{CADP\ –\ pomiar\ pierwszy} – CT_{CADP\ –\ pomiar\ drugi}$]; różnica [$CT_{CEPI\ –\ pomiar\ pierwszy} – CT_{CEPI\ –\ pomiar\ drugi}$]) względem średnich ($\bar{x}$) wyników dla obu pomiarów dla każdej kasety (średnia [$CT_{CADP\ –\ pomiar\ pierwszy}$; $CT_{CADP\ –\ pomiar\ drugi}$]; średnia [$CT_{CEPI\ –\ pomiar\ pierwszy}$; $CT_{CEPI\ –\ pomiar\ drugi}$]).

Dla kaset zawierających kolagen i ADP (CADP) mamy:



Dla kaset zawierających kolagen i epinefrynę (CEPI) uzyskujemy:



Obie pary wykresów pokazują, że nie występuje zależność między wartościami różnic a wartościami średnich, zarówno dla kaset CADP, jak i dla kaset CEPI, ale wśród pomiarów w kasetach z kolagenem i epinefryną występują odstające obserwacje, mieszczące się poza

zakresem średniej ±2SD. Możemy przypuszczać, że zostały one spowodowane najprawdopodobniej przez błąd techniczny podczas wykonywania pomiaru lub czynnik losowy (na przykład przeterminowane lub niewłaściwie przechowywane kasety pomiarowe). Ponieważ te odstające pomiary są pojedyncze (2 na 30 obserwacji co stanowi 7% wszystkich przypadków), a zatem są raczej czymś wyjątkowym niż regułą, możemy przypadków tych nie uwzględniać w dalszej części analizy.

W poniższej tabeli zestawiono różnice oraz ich kwadraty dla poszczególnych par pomiarów zebranych przy użyciu obu typów kaset.

dawca	CT_{CADP} (sek.)		CT_{CEPI} (sek.)		różnica (pomiar I – pomiar II) (sek.)		kwadrat różnicy (pomiar I – pomiar II) (sek.^{2})	
	pomiar pierwszy	pomiar drugi	pomiar pierwszy	pomiar drugi	CT_{CADP}	CT_{CEPI}	CT_{CADP}	CT_{CEPI}
1	140.0	151.0	148.0	140.0	–11.0	8.0	121.0	64.0
2	126.0	106.0	79.0	88.0	20.0	–9.0	400.0	81.0
3	144.0	134.0	116.0	104.0	10.0	12.0	100.0	144.0
4	69.0	69.0	118.0	107.0	0.0	11.0	0.0	121.0
5	69.0	61.0	82.0	85.0	8.0	–3.0	64.0	9.0
6	81.0	86.0	120.0	114.0	–5.0	6.0	25.0	36.0
7	105.0	83.0	110.0	118.0	22.0	–8.0	484.0	64.0
8	103.0	96.0	85.0	79.0	7.0	6.0	49.0	36.0
9	114.0	143.0	110.0	118.0	–29.0	–8.0	841.0	64.0
10	77.0	70.0	100.0	115.0	7.0	–15.0	49.0	225.0
11	80.0	81.0	158.0	182.0	–1.0	–24.0	1.0	576.0
12	80.0	72.0	142.0	181.0	8.0	–39.0	64.0	1521.0
13	49.0	68.0	173.0	182.0	–19.0	–9.0	361.0	81.0
14	55.0	77.0	161.0	165.0	–22.0	–4.0	484.0	16.0
15	80.0	80.0	74.0	65.0	0.0	9.0	0.0	81.0
16	72.0	68.0	54.0	112.0	4.0	–58.0	16.0	3364.0
17	86.0	77.0	77.0	86.0	9.0	–9.0	81.0	81.0
18	73.0	66.0	131.0	141.0	7.0	–10.0	49.0	100.0
19	146.0	106.0	89.0	96.0	40.0	–7.0	1600.0	49.0
20	75.0	82.0	116.0	132.0	–7.0	–16.0	49.0	256.0
21	51.0	49.0	132.0	158.0	2.0	–26.0	4.0	676.0
22	96.0	72.0	100.0	109.0	24.0	–9.0	576.0	81.0
23	92.0	59.0	127.0	145.0	33.0	–18.0	1089.0	324.0
24	62.0	51.0	173.0	182.0	11.0	–9.0	121.0	81.0
25	54.0	48.0	161.0	165.0	6.0	–4.0	36.0	16.0
26	82.0	120.0	74.0	65.0	–38.0	9.0	1444.0	81.0
27	73.0	64.0	110.0	54.0	9.0	56.0	81.0	3136.0
28	75.0	84.0	77.0	86.0	–9.0	–9.0	81.0	81.0
29	95.0	125.0	131.0	141.0	–30.0	–10.0	900.0	100.0
30	77.0	62.0	82.0	85.0	15.0	–3.0	225.0	9.0
średnia	86.0	83.7	113.7	120.0	2.4	–6.3	313.2	385.1
odch. std.	26.2	27.5	32.7	38.0	17.8	18.9	440.5	833.6
suma	2581.0	2510.0	3410	3600.0	71.0	–190.0	**9395**	**11554.0**

Obliczamy współczynniki powtarzalności dla obu typów kaset pomiarowych. Dla kaset CADP mamy:

$$s_{rep} = 2s_d = 2\sqrt{\frac{\sum_{i=1}^{n} d^2}{n}} = 2\sqrt{\frac{9395}{30}} = 35.4 \text{ sek.}$$

Dla kaset CEPI analogicznie obliczona wartość współczynnika powtarzalności wynosi 39.2 sek.

Rozdział 18

Określanie niezbędnej liczebności próby

Przykład 53

W grupie pacjentów zamierzamy przebadać skuteczności dwóch leków: A i B. Zakładamy, że skuteczność leku A ma być przynajmniej w 70% przypadków większa niż skuteczność leku B, tzn. u 7 na 10 pacjentów lek A odniesie bardziej pozytywny skutek, podczas gdy w przypadku leku B proporcja ta będzie wynosić 5 na 10. Chcemy osiągnąć wynik istotny statystycznie na poziomie istotności nie mniejszym niż 5% przy mocy wnioskowania 90%. Ilu pacjentów powinniśmy przebadać?

Załóżmy, że przebadamy grupę 20-osobową. Wynik będzie istotny na poziomie 5%, jeżeli będzie oddalony przynajmniej o 1.96 wielokrotności błędu standardowego od średniej. Nasza hipoteza zerowa zakłada, że wartość proporcji zarówno w przypadku leku A, jak i leku B wynosi 0.5 (tzn. u połowy pacjentów lek ten odniesie pozytywny skutek, a u drugiej połowy nie). Odpowiednio, błąd standardowy wynosi $\sqrt{(0.5 \times 0.5 / n)}$. Dla próby o liczebności 20:

$$SE = \sqrt{(0.5 \times 0.5 / 20)} = 0.1118$$

$$0.5 + 1.96 \text{ x SE} = 0.5 + 1.96 \text{ x } 0.1118 = 0.72$$

$$\text{oraz } 0.5 - 1.96 \text{ x SE} = 0.5 - 1.96 \text{ x } 0.1118 = 0.28$$

Zatem wartość, która byłaby istotnie statystycznie różna od 0.5 musiałaby leżeć powyżej 0.72 lub poniżej 0.28. Skoro rzeczywista proporcja wynosi 0.7, jakie jest prawdopodobieństwo, że zebrane obserwacje dadzą wartość powyżej 0.72 i istotny wynik testu? Prawdopodobieństwo to odpowiada szaremu polu pod krzywą, dla której wartość średnia wynosi 0.7 zaś SE równa się $\sqrt{(0.7 \times 0.3 / 20)} = 0.1025$ (Ryc. 18).

Obliczona wartość *z* wynosi:

$$\frac{0.72 - 0.7}{0.1025} = 0.20$$

a wartość dystrybuanty 0.421. Możemy zatem powiedzieć, że w próbie obejmującej 20 pacjentów mamy jedynie 42.1% szans na wykazanie, że lek A wpływa ze skutecznością 70% na poprawę stanu zdrowia pacjentów.

Jeżeli zwiększymy próbę do 50 pacjentów, to uzyskamy wyniki:

$$SE = \sqrt{(0.5 \times 0.5 / 50)} = 0.07071$$

$$0.5 + 1.96 \text{ x SE} = 0.5 + 1.96 \text{ x } 0.07071 = 0.64$$

$$\text{oraz } 0.5 - 1.96 \text{ x SE} = 0.5 - 1.96 \text{ x } 0.07071 = 0.36$$

Dla SE $\sqrt{(0.7 \times 0.3 / 50)} = 0.0648$ (Ryc. 7) obliczona wartość *z* wynosi:

$$\frac{0.64 - 0.7}{0.0648} = -0.926$$

a wartość dystrybuanty $1 - 0.1762 = 0.8238$. Zatem w próbie obejmującej 50 pacjentów istnieje 82.4% szans na wykazanie, że lek A wpływa ze skutecznością 70% na poprawę stanu zdrowia pacjentów (Ryc. 18).

Aby osiągnąć prawdopodobieństwo 90% orzekania istotności różnicy w stosunku do proporcji 0.5, należy jeszcze powiększyć grupę. O ile? Wynik będzie istotny powyżej wartości 0.5 + 1.96 x SE lub poniżej wartości 0.5 – 1.96 x SE.

Potrzebujemy próby wystarczająco licznej, aby 90% rozkładu przypadało powyżej punktu wyznaczonego przez wartość 0.5 + 1.96 x SE. Wartość z dla próby badanej odpowiadająca 90% wynosi –1.28. Odpowiada to wartości obserwowanej:

$$0.7 - 1.28 \times SE = 0.7 - 1.28 \times \sqrt{(0.7 \times 0.3 / n)}$$

Zatem liczebność próby *n* powinna być wystarczająco duża aby

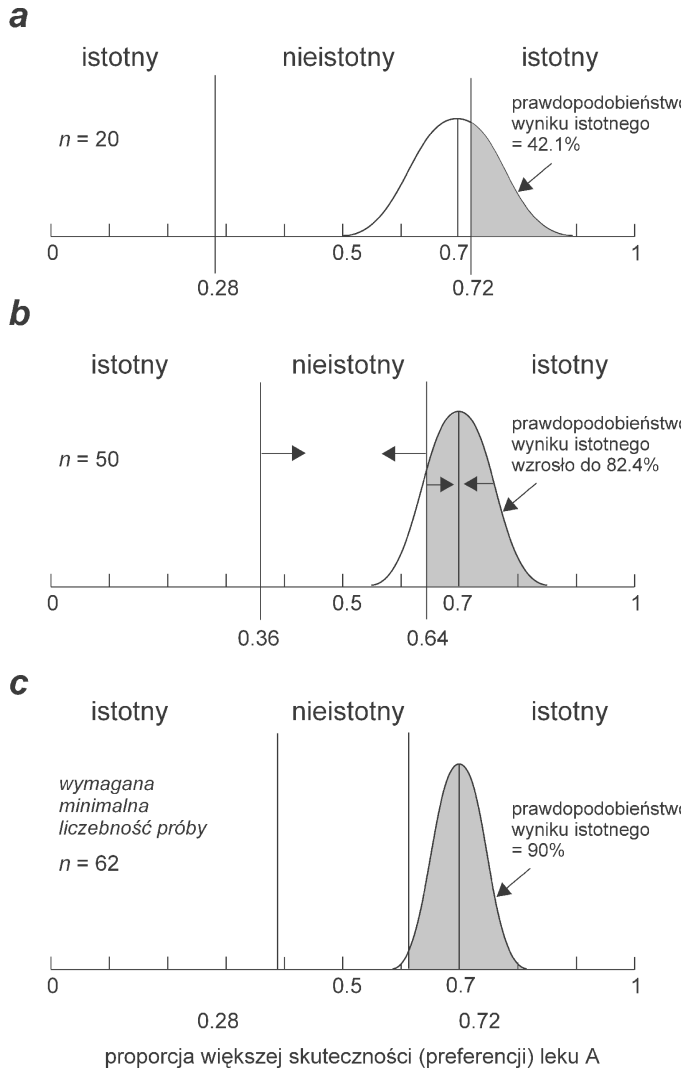
$$0.7 - 1.28 \times \sqrt{(0.7 \times 0.3 / n} \; > \; 0.5 + 1.96 \times \sqrt{(0.5 \times 0.5 / n)}$$

Po przekształceniu otrzymujemy

$$0.7 - 0.5 > \frac{1.96 \times \sqrt{(0.5 \times 0.5)} + 1.28 \times \sqrt{(0.7 \times 0.3)}}{\sqrt{n}}$$

$$n > \frac{[1.96 \times \sqrt{(0.5 \times 0.5)} + 1.28 \times \sqrt{(0.7 \times 0.3)}]^2}{(0.2)^2}$$

$$n > \frac{1.5666^2}{(0.2)^2} = 61.4$$



Ryc. 18. Prawdopodobieństwo uzyskania 5% istotności różnic przy różnych liczebnościach (*n*) próby przy testowaniu hipotezy o skuteczności leku A większej w porównaniu z lekiem B w 70% przypadków (proporcja 0.7), podczas gdy hipoteza zerowa zakłada proporcję 0.5.

Nasza próba powinna liczyć 62 pacjentów, aby z prawdopodobieństwem 90% wykazać istotną różnicę w skuteczności między lekiem A i lekiem B, jeżeli założymy, że lek A ma być przynajmniej w 70% przypadków bardziej skuteczny niż lek B.

Stosując wzór na obliczanie liczebności próby dla pojedynczej proporcji (zobacz Tab. 4 w Rozdziale *„Metody estymacji liczebności próby"*):

$$n > \frac{\{u\sqrt{[\pi(1-\pi)]} + v\sqrt{[\pi_0(1-\pi_0)]}\}^2}{(\pi-\pi_0)^2}$$

gdzie:
n minimalna wymagana liczebność próby
π proporcja próby badanej (dla hipotezy alternatywnej)
π_0 proporcja teoretyczna (dla hipotezy zerowej)
u punkt krytyczny jednostronny rozkładu normalnego odpowiadający proporcji 100% – moc testu, tzn. jeżeli moc wynosi na przykład 90%, to (100% – 90%) = 10% i $u = 1.28$
v punkt krytyczny rozkładu normalnego odpowiadający wymaganemu (obustronnemu) poziomowi istotności np. dla istotności 5%, $v = 1.96$,

Dla powyższego przykładu
$\pi = 0.7 \quad \pi_0 = 0.5 \quad u = 1.28 \quad \text{i} \quad v = 1.96$

$$n > \frac{[1.28 \times \sqrt{(0.7 \times 0.3)} + 1.96 \times \sqrt{(0.5 \times 0.5)}]^2}{(0.7 - 0.5)^2} \qquad n > \frac{2.4542}{0.04} = 61.4$$

Przykład 54

Chcemy zbadać czy suplementacja witamin u kobiet w ciąży wpływa na zwiększenie masy urodzeniowej noworodków. Kobiety ciężarne zgłaszające się do poradni prenatalnej są w sposób przypadkowy przydzielane do grupy otrzymującej suplementację lub nie. Ile kobiet powinna obejmować każda grupa?

Założenia:
a) zdecydowano, że minimalna różnica mas urodzeniowych powinna wynosić 0.25 kg, czyli $\mu_1 - \mu_2 = 0.25$ kg,
b) na podstawie kilkuletnich obserwacji oszacowano, że odchylenie standardowe mas urodzeniowych dzieci matek zgłaszających się do poradni wynosi około 0.4 kg; zatem przyjęto, że $\sigma_1 = 0.4$ kg i $\sigma_2 = 0.4$ kg,
c) zdecydowano, że moc wnioskowania będzie wynosiła 95%; odpowiednia wartość $u = 1.64$,
d) przyjmujemy, że chcemy uzyskać wynik istotny na poziomie 1% (0.01), czyli wartość dystrybuanty (obustronnej) wynosi $v = 2.58$.
Stosując równanie (4) z Tabeli 4 (str. 94):

$$n > \frac{(u+v)^2(\sigma_1^2 + \sigma_2^2)}{(\mu_1 - \mu_2)^2} \qquad n > \frac{(1.64 + 2.58)^2(0.4^2 + 0.4^2)}{(0.25)^2} \qquad n > \frac{17.8084 \times 0.32}{0.0625} = 91.2$$

Badaniem powinniśmy zatem objąć przynajmniej po 90 kobiet w każdej z grup.

Przykład 55

Pragniemy ustalić jaka jest częstość występowania ostrych biegunek u dzieci w wieku do 5 lat zamieszkujących tereny wiejskie. Szacujemy, że częstość taka wynosi średnio około 3 epizodów u jednego dziecka w ciągu jednego roku, ale chcemy dokonać oceny z dokładnością ±0.2, to znaczy jeżeli zaobserwujemy średnio 3 epizody to chcielibyśmy wniosko-

wać, że częstość rzeczywista waha się między 2.8 a 3.2. Pragniemy zatem, aby nasz 95% przedział ufności nie był większy niż ±0.2. Oznacza to, że nasz błąd standardowy będzie wynosił około 0.1 epizodu/dziecko/rok (ponieważ wiemy, że $CI = \bar{x} \pm 1.96 \times SE$). Stosując wzór (9) z tabeli 4 (str. 96).

$$n > \frac{\mu}{e^2} \quad \text{otrzymujemy} \quad n > \frac{3}{(0.1)^2} = 300$$

Czyli potrzebujemy obserwować każde z 300 dzieci przez jeden rok, lub odpowiednio czterokrotnie więcej, tzn. każde z 1200 dzieci w ciągu 3 miesięcy. W badaniach tego typu należy zawsze zwracać uwagę na wpływ sezonowości na obserwowaną częstość.

Przykład 56

Planujemy przeprowadzić badania populacyjne typu *case-control* dotyczące wpływu karmienia piersią (grupa kontrolna) lub butelką (grupa badana) na śmiertelność z powodu infekcji układu oddechowego. Oczekujemy, że około 40% dzieci z grupy kontrolnej ($\pi_1 = 0.4$) będzie karmionych butelką, oraz chcielibyśmy wykryć różnicę związaną z dwukrotnym podwyższeniem ryzyka zgonu u takich dzieci w stosunku do dzieci karmionych piersią (OR = 2.0). Jak wiele przypadków kontrolnych i badanych musimy włączyć do badania, jeżeli chcielibyśmy wnioskować z mocą nie mniejszą niż 90% (u = 1.28) przy poziomie istotności różnic 5% (v = 1.96)?

Korzystamy z równania (7) w Tabeli 4 (str. 95). Proporcja przypadków badanych wynosi:

$$\pi_2 = \frac{\pi_1 \times OR}{1 + \pi_1 (OR - 1)} = \frac{0.4 \times 2}{1 + 0.4 \times (2 - 1)} = \frac{0.8}{1.4} = 0.57$$

Na podstawie π_1 i π_2 liczymy

$$\bar{\pi} = \frac{0.4 + 0.57}{2} = 0.485$$

Obliczamy minimalną liczebność próby na podstawie równania (7):

$$n > \frac{\{u\sqrt{[\pi_1(1-\pi_1) + \pi_2(1-\pi_2)]} + v\sqrt{[2\bar{\pi}(1-\bar{\pi})]}\}^2}{(\pi_2 - \pi_1)^2}$$

$$n > \frac{[1.28\sqrt{(0.4 \times 0.6 + 0.57 \times 0.43)} + 1.96\sqrt{(2 \times 0.485 \times 0.515)}]^2}{(0.57 - 0.4)^2}$$

$$n > \frac{[1.28\sqrt{0.4851} + 1.96\sqrt{0.4996}]^2}{(0.17)^2} = \frac{2.2769^2}{(0.17)^2} = 179.4$$

Potrzebujemy zbadać po 180 przypadków kontrolnych i badanych, razem – 360 osób. Jeżeli zdecydowalibyśmy się na przebadanie 3-krotnie więcej dzieci kontrolnych (na przykład dlatego, że zebranie grupy kontrolnej może być w praktyce o wiele łatwiejsze niż przypadków chorobowych), to zgodnie z Tabelą 8 obliczymy, że dla $c = 3$ odpowiedni współczynnik dopasowania liczebności wynosi 2/3. Oznacza to, że będziemy potrzebowali 180 x 2/3 = 120 przypadków badanych i 3-krotnie tyle kontroli, czyli 360 osób. Jak widać, chociaż liczebność dla grupy badanej obniżyła się półtora raza, to całkowita liczebność wzrosła z 360 do 480.

Tab. 8. Współczynniki dopasowania liczebności przy porównywaniu grup o nierównych liczebnościach, przy doborze wielokrotnych kontroli dla każdego badanego przypadku. Współczynnik (*f*) odnosi się do mniejszej grupy i wynosi $(c+1)/(2c)$, gdzie c oznacza wielokrotność liczebności większej grupy względem mniejszej. Jeżeli estymowana liczebność wynosi n, to liczebność mniejszej grupy wynosi fn, zaś większej odpowiednio cfn.

ile razy liczebność większej grupy przewyższa liczebność mniejszej grupy	**współczynnik korekcji liczebności mniejszej grupy (*f*)**
1	1
2	3/4
3	2/3
4	5/8
5	3/5
6	7/12
7	4/7
8	9/16
9	5/9
10	11/20
20	21/40

Przykład 57

Dla danych przykładu 24 chcemy obliczyć, jak duża powinna być próba badana, abyśmy mogli odrzucić hipotezę zerową mówiącą, że intensywny wysiłek fizyczny wpływa na zmianę masy ciała szczurów? Testowanie planujemy przeprowadzić przy poziomie istotności 0.05 oraz z 90% prawdopodobieństwem wykrycia różnicy różnej od zera przynajmniej o 1.0 g.

Zmienność próby wynosiła $s = 1.2523$ g, czyli $s^2 = 1.568$ g^2. Różnica, jaką chcemy wykryć ma wynosić 1 g, czyli $\delta = \mu - \mu_0 = 1.0$ g.

Ponieważ nasza próba jest próbą losową o skończonej liczebności, wylosowaną w sposób przypadkowy z populacji wszystkich szczurów, nie możemy mieć pewności, czy parametry rozkładu naszej próby aproksymują chociaż w przybliżonym stopniu do parametrów rzeczywistych ogólnej populacji, μ oraz σ. Dlatego przyjmujemy następujące wartości krytyczne: $v = t_{\alpha(2),\, n-1}$ oraz $u = t_{\beta(1),\, n-1}$. W przypadku, gdybyśmy znali rzeczywiste σ, zamiast t_α oraz t_β zastosowalibyśmy z_α i z_β. Nie znamy liczby stopni swobody, ponieważ zależy ona od liczebności, której oczywiście także nie znamy, i którą dopiero chcemy obliczyć. Niedoświadczonemu badaczowi może się wydawać, że to rodzaj „błędnego koła", którego nie sposób rozwikłać. W praktyce, w sytuacjach kiedy liczba stopni swobody jest nieznana, jest ona szacowana *a priori* metodami iteracji (czyli kolejnych dopasowań; niemal zawsze iterację taką „powierza się" odpowiednim programom komputerowym). W naszym

konkretnym przypadku, załóżmy, że należałoby przebadać 20 zwierząt. Wtedy, jeżeli moc testu ma być 90%, czyli $\beta = 1 - 0.9 = 0.1$, wartość krytyczna będzie wynosić $t_{0.10(1),19} = u = 1.33$. Odpowiednio, $v = t_{\alpha(2),\, n-1} = 2.09$.

Szacowana liczebność próby wynosi:

$$n = \frac{(u+v)^2 \sigma^2}{(\mu - \mu_0)^2}$$

$$n = \frac{1.5682 \times (2.09 + 1.33)^2}{(1.0)^2} = 18.34$$

czyli próba powinna liczyć 18-19 obserwacji. Analogiczna wartość szacunkowa dla założonej *a priori* liczebności $n = 19$ szczurów (z $v = 2.20$ i $u = 1.33$) wynosiłaby $n = 18.5$. Możemy zatem wnioskować, że próba badana powinna liczyć co najmniej 19 obserwacji.

Przykład 58

Jeżeli zastosujemy test obustronny do wyników z przykładu 24, to jaka będzie najmniejsza wykrywana różnica (tzn. różnica między μ i μ_0), którą możemy wykryć z mocą 90% przy poziomie istotności 0.05 w próbie liczącej 25 obserwacji?

Przekształcając równanie:

$$n = \frac{(u+v)^2 \sigma^2}{(\mu - \mu_0)^2}$$

uzyskujemy:

$$\mu - \mu_0 = \sqrt{\frac{\sigma^2}{n}}(u+v)$$

Dla próby losowej o niewielkiej liczebności możemy zapisać:

$$\delta = \sqrt{\frac{s^2}{n}}(t_{\alpha,n-1} + t_{\beta(1),n-1}),$$

czyli

$$\delta = \sqrt{\frac{1.5682}{12}}(2.06 + 1.32) = (0.2505)(3.38) = 0.845$$

Najmniejsza różnica jaką będziemy mogli wykryć wynosi 0.845 g.

Przykład 59

Dla danych z tego samego przykładu chcemy ustalić prawdopodobieństwo poprawnego odrzucenia fałszywej hipotezy zerowej, czyli chcemy znać moc testu. Przekształcamy równanie:

$$\delta = \sqrt{\frac{s^2}{n}}(t_{\alpha,n-1} + t_{\beta(1),n-1})$$

$$t_{\beta(1),n-1} = \frac{\delta}{\sqrt{\frac{s^2}{n}}} - t_{\alpha,n-1}$$

Dla $n - 1 = 11$, $t_{0.05(2),\, n-1} = 2.201$ i $s^2 = 1.5682$ uzyskujemy:

$$t_{\beta(1),n-1} = \frac{1.0}{\sqrt{\frac{1.5682}{12}}} - 2.201 = 0.57$$

Z tablic rozkładu t Studenta odczytujemy, że dla $n - 1 = 11$ stopni swobody $\beta > 0.25$, czyli moc wynosi $1 - \beta < 0.75$. Jeżeli założymy, że nasze , $t_{\beta(1),n-1} \approx z_{\beta(1),n-1}$ to analogiczna wartość krytyczna dla rozkładu normalnego wynosi $\beta = 0.284$, czyli moc testu równa się około 72%.

Rozdział 19

Zastosowanie metod transformacji danych do „normalizacji" rozkładu

Przykład 60

Badano stan zakażenia pasożytami przewodu pokarmowego u bydła, na podstawie oceny liczby jaj pasożytów znajdowanych w odchodach zwierząt. Należy obliczyć średnią liczbę jaj pasożytów wykrywanych w odchodach pojedynczego zwierzęcia.

Dane, uporządkowane rosnąco, przedstawiają się następująco:

lp	x	x+1	log(x+1)	lp	x	x+1	log(x+1)
1	0	1	0.00	46	13	14	1.15
2	0	1	0.00	47	14	15	1.18
3	0	1	0.00	48	14	15	1.18
4	0	1	0.00	49	15	16	1.20
5	0	1	0.00	50	15	16	1.20
6	0	1	0.00	51	16	17	1.23
7	1	2	0.30	52	17	18	1.26
8	1	2	0.30	53	17	18	1.26
9	1	2	0.30	54	18	19	1.28
10	1	2	0.30	55	21	22	1.34
11	1	2	0.30	56	28	29	1.46
12	1	2	0.30	57	28	29	1.46
13	1	2	0.30	58	29	30	1.48
14	2	3	0.48	59	31	32	1.51
15	2	3	0.48	60	35	36	1.56
16	2	3	0.48	61	36	37	1.57
17	2	3	0.48	62	39	40	1.60
18	3	4	0.60	63	41	42	1.62
19	3	4	0.60	64	43	44	1.64
20	3	4	0.60	65	47	48	1.68
21	4	5	0.70	66	49	50	1.70
22	4	5	0.70	67	51	52	1.72
23	4	5	0.70	68	57	58	1.76
24	4	5	0.70	69	57	58	1.76
25	4	5	0.70	70	67	68	1.83
26	4	5	0.70	71	68	69	1.84
27	4	5	0.70	72	73	74	1.87
28	5	6	0.78	73	73	74	1.87
29	5	6	0.78	74	74	75	1.88
30	6	7	0.85	75	80	81	1.91

31	6	7	0.85	76	124	125	2.10
32	7	8	0.90	77	129	130	2.11
33	8	9	0.95	78	158	159	2.20
34	8	9	0.95	79	173	174	2.24
35	9	10	1.00	80	189	190	2.28
36	9	10	1.00	81	201	202	2.31
37	10	11	1.04	82	220	221	2.34
38	10	11	1.04	83	278	279	2.45
39	10	11	1.04	84	280	281	2.45
40	11	12	1.08	85	459	460	2.66
41	11	12	1.08	86	541	542	2.73
42	12	13	1.11	87	767	768	2.89
43	12	13	1.11	88	830	831	2.92
44	12	13	1.11	89	938	939	2.97
45	13	14	1.15				

Jak można zauważyć, rozkład zebranych obserwacji (kolumna „*x*" w tabeli z wynikami) jest wyraźnie prawoskośny (Ryc. 19). Dla takiego wyraźnie asymetrycznego rozkładu średnia arytmetyczna oraz odchylenie standardowe nie będą reprezentatywne. Obliczmy wartości średniej arytmetycznej i mediany. Wynoszą one:

$$\bar{x}_a = 74.9 \quad Me = 13.0$$

Te bardzo różne wartości średniej arytmetycznej oraz mediany są potwierdzeniem prawoskośności rozkładu. Wartość odchylenia średniej arytmetycznej wynosi: $s = 171.5$, czyli ponad dwukrotnie przewyższa średnią; współczynnik zmienności dla tej pary parametrów wyniósłby CV = 100% x (74.9/171.5) = 229%.

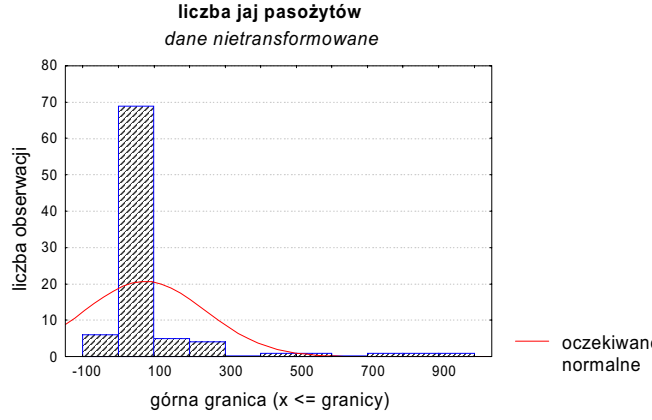
Z uwagi na asymetryczność rozkładu nie możemy podać obiektywnie przedziału ufności dla średniej. Skoro mamy do czynienia z rozkładem prawoskośnym, nasze dane liczby jaj pasożytów moglibyśmy poddać transformacji logarytmicznej, ale dla licznych obserwacji zarejestrowano wartości równe zero ($\log(0) = -\infty$). Problem ten możemy łatwo ominąć dodając do każdego wyniku liczbę 1 (*kolumna „x+1" w tabeli z wynikami*). Rozkład tak zmodyfikowanych danych jest nadal prawoskośny (Ryc. 19a-b), ale można go łatwo poddać transformacji $\log_{10}$ (*kolumna „log(x+1)" w tabeli wyników*), uzyskując rozkład lognormalny (Ryc. 19a-c). Dla takiej zbiorowości wyników obliczamy średnią geometryczną, która wynosi:

$$\bar{x}_g = antilog\left(\frac{109.18}{89}\right) - 1 = antilog(1.2267) - 1 = 10^{1.2267} - 1 = 16.855 - 1 = 15.855$$

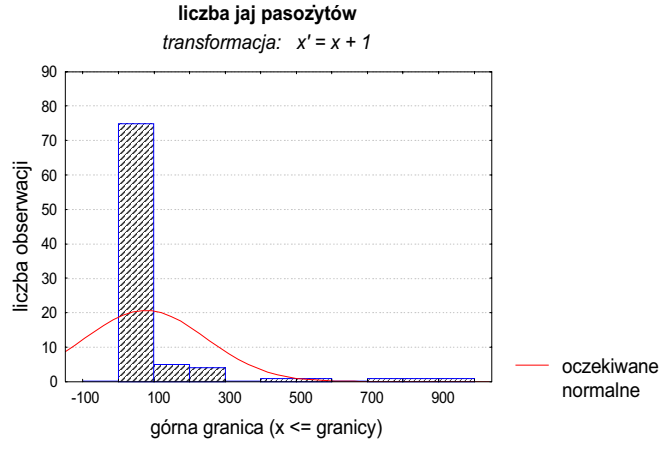
(*Ponieważ wartość średniej geometrycznej obliczaliśmy w oparciu o logarytmowane dane powiększone o 1 – w celu ominięcia problemu liczenia log(0) – to końcowy wynik musimy także pomniejszyć o 1*).

Widzimy, że średnia geometryczna jest zbliżona do wartości mediany i znacznie lepiej charakteryzuje tendencję centralną liczby jaj pasożytów niż średnia arytmetyczna.

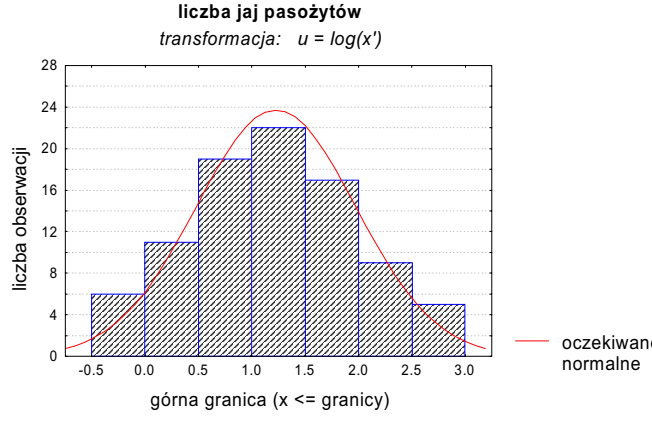
a)



b)



c)



Ryc. 19. Rozkład nietransformowanych danych liczby jaj pasożytów przed (a) i po modyfikacji matematycznej (b) jest wyraźnie prawoskośny. Transformacja logarytmiczna przywraca symetryczność rozkładu lognormalnego (c).

Przykład 61

Płytki krwi pobranej na wersenian od zdrowych ochotników oraz pacjentów z cukrzycą typu 2 inkubowano w ciągu pół godziny w temperaturze pokojowej. Oceniano stopień aktywacji płytek krwi na podstawie ekspresji selektyny P uwalnianej z ziarnistości α płytek. Czy płytki osób z cukrzycą uwalniają więcej selektyny P (aktywują się bardziej) niż płytki zdrowych ochotników?

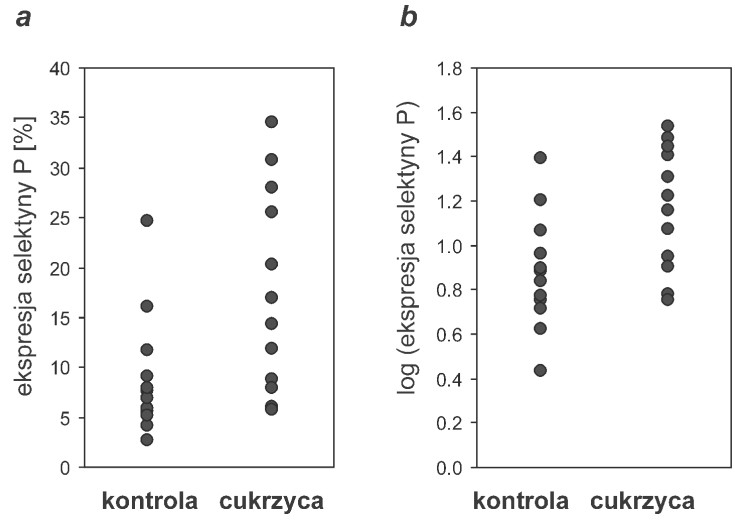
	ekspresja selektyny P (%)		**log (ekspresja selektyny P)**	
	zdrowi	**cukrzyca**	**zdrowi**	**cukrzyca**
	16.2	25.7	1.21	1.41
	7.7	12.0	0.88	1.08
	9.2	14.4	0.96	1.16
	11.7	17.0	1.07	1.23
	5.7	30.9	0.75	1.49
	5.9	6.1	0.77	0.78
	24.8	8.9	1.39	0.95
	6.9	8.1	0.84	0.91
	2.7	28.1	0.44	1.45
	4.2	5.8	0.62	0.76
	5.2	34.6	0.72	1.54
	8.0	20.4	0.90	1.31
średnia	9.0	17.6	0.88	1.17
SD	6.1	10.1	0.26	0.28
n	12	12	12	12

Rycina 20 pokazuje, że rozkłady obserwacji zebranych w obu grupach są wyraźnie prawoskośne (liczne obserwacje o wysokich i bardzo wysokich wartościach peryferycznych), zaś na podstawie powyższej tabeli widać, że odchylenie dla grupy osób z cukrzycą jest prawie dwukrotnie wyższe niż w grupie osób zdrowych. W celu „normalizacji" rozkładów użyjemy transformacji logarytmicznej (*prawe kolumny tabeli* i Ryc. 20). Z tabeli widzimy, że wartości odchyleń logarytmowanych danych są niemal równe, a rozkłady danych (Ryc. 20) uległy symetryzacji. Ponieważ dane logarytmowane lepiej spełniają warunki stosowania testów parametrycznych, dlatego do porównania dwóch średnich użyjemy parametrów obliczonych dla zbiorów danych transformowanych.
Nasze hipotezy maja postać:
H_0: $\mu_{cuk} \leq \mu_{zdr}$
H_A: $\mu_{cuk} > \mu_{zdr}$

$$s = \sqrt{[(11 \times 0.26^2 + 11 \times 0.28^2)/22]} = 0.27$$

$$t = \frac{1.17 - 0.88}{0.27\sqrt{(1/12 + 1/12)}} = 6.44 \qquad \text{dla } d.f. = 12+12-2 = 22 \text{ stopni swobody}$$

Istotność różnic sprawdzimy dla prawdopodobieństwa 99.9% czyli nasz poziom istotności będzie wynosić 0.001. Wartość krytyczna odczytana z tablic testu jednostronnego (pytanie brzmiało: *„czy ekspresja jest wyższa w cukrzycy niż w kontroli*?') dla 22 stopni



Ryc. 20. Ekspresja selektyny P w płytkach zdrowych ochotników oraz osób z cukrzycą typu 2: (a) dane nietransformowane, (b) dane logarytmowane ($\log_{10}$).

swobody i poziomu istotności $\alpha = 0.001$ wynosi $t_{0.001(1),22} = 3.50$ i jest niższa od obliczonej wartości $t_{dośw}$. Zatem z prawdopodobieństwem nie mniejszym niż 0.1% możemy orzec, że ekspresja selektyny P w płytkach osób z cukrzycą inkubowanych w ciągu 30 minut w temperaturze pokojowej jest wyższa niż u osób zdrowych.

Obliczmy jeszcze średnią geometryczną, odchylenia standardowe średniej geometrycznej oraz przedziały ufności. Dla grupy zdrowych osób średnia geometryczna, jej odchylenie i granice przedziału ufności wynoszą:

$$\bar{x}_g = antilog(0.88) = 10^{0.88} = 7.6\ \% \quad SD_g = antilog(0.26) = 10^{0.26} = 1.8\ \%$$

$$CI(95\%) = 0.88 \pm 2.20 \times 0.26\sqrt{12} = \text{od } 0.715 \text{ do } 1.045$$

(2.20 jest wartością punktu krytycznego dla poziomu istotności 0.05 przy 11 stopniach swobody)
czyli po odlogarytmowaniu:

$$\text{od } antilog(0.715) = 10^{0.715} = 5.2\% \qquad \text{do } antilog(1.045) = 10^{1.045} = 11.1\%$$

Dla grupy osób z cukrzycą:

$$\bar{x}_g = antilog(1.17) = 10^{1.17} = 14.8\ \% \quad SD_g = antilog(0.28) = 10^{0.28} = 1.9\ \%$$

$$CI(95\%) = 1.17 \pm 2.20 \times 0.28\sqrt{12} = \text{od } 0.992 \text{ do } 1.34\%$$

czyli po odlogarytmowaniu:

$$\text{od } antilog(0.992) = 10^{0.992} = 9.8\% \qquad \text{do } antilog(1.348) = 10^{1.348} = 22.3\%$$

Średnią geometryczną, odchylenia standardowe średniej geometrycznej oraz przedziały ufności zestawiono w tabeli:

	średnia geometryczna	**odchylenie średniej geometrycznej**	**granice 95% przedziału ufności**	
osoby zdrowe	7.6	1.8	5.2	11.1
osoby z cukrzycą	14.8	1.9	9.8	22.3

Zwróćmy uwagę na następującą prawidłowość: przedział ufności nie jest symetryczny względem średniej geometrycznej, ale iloraz górnej granicy przedziału do średniej 11.1/7.6 = 1.46 jest taki sam jak iloraz średniej do dolnej granicy przedziału 7.6/5.2 = 1.46. Jest tak dlatego, że odchylenie standardowe na skali logarytmicznej jest multiplikatywną a nie addytywną miarą rozrzutu wokół średniej.

Rozdział 20

Metody badania zależności statystycznych między zmiennymi

Korelacja liniowa (Pearsona) i nieliniowa (eta)

Przykład 62

W tabeli poniżej przedstawiono wyniki ekspresji antygenów powierzchniowych w błonach płytek krwi rozpoznawanych przez przeciwciała monoklonalne anty-CD62 i PAC-1 dla płytek w stanie spoczynku oraz po aktywacji kolagenem. Jaki jest związek między ekspresją jednego i drugiego antygenu?

	CD62		PAC-1	
dawca	**spoczynkowe (x_1)**	**aktywowane (x_2)**	**spoczynkowe (y_1)**	**aktywowane (y_2)**
1	2.0	79.8	1.7	77.0
2	5.2	67.2	0.6	41.7
3	1.8	61.3	1.3	60.7
4	2.0	71.4	6.4	71.1
5	3.7	53.6	5.9	55.9
6	1.4	38.8	3.7	50.0
7	6.3	48.5	2.5	59.3
8	2.4	76.6	2.2	57.6
9	1.9	45.6	1.3	58.8
10	1.7	51.8	1.1	62.4
11	1.6	62.4	1.4	66.2
12	1.8	18.2	1.7	17.4
13	3.6	33.4	1.7	30.1
14	4.5	74.1	3.9	61.5
15	1.0	42.6	1.7	62.4
$\overline{X}_i$	2.7	55.0	2.5	55.5
$\sum X_i$	40.9	825.3	37.1	832.1

Obliczenia:

x_1^2	y_1^2	x_2^2	y_2^2	$x_1 y_1$	$x_2 y_2$
4.0	2.9	6368.0	5929.0	3.4	6144.6
27.0	0.4	4515.8	1738.9	3.1	2802.2
3.2	1.7	3757.7	3684.5	2.3	3720.9

4.0	41.0	5098.0	5055.2	12.8	5076.5
13.7	34.8	2873.0	3124.8	21.8	2996.2
2.0	13.7	1505.4	2500.0	5.2	1940.0
39.7	6.3	2352.3	3516.5	15.8	2876.1
5.8	4.8	5867.6	3317.8	5.3	4412.2
3.6	1.7	2079.4	3457.4	2.5	2681.3
2.9	1.2	2683.2	3893.8	1.9	3232.3
2.6	2.0	3893.8	4382.4	2.2	4130.9
3.2	2.9	331.2	302.8	3.1	316.7
13.0	2.9	1115.6	906.0	6.1	1005.3
20.3	15.2	5490.8	3782.3	17.6	4557.2
1.0	2.9	1814.8	3893.8	1.7	2658.2
$\sum x_1^2$	$\sum y_1^2$	$\sum x_2^2$	$\sum y_2^2$	$\sum x_1 y_1$	$\sum x_2 y_2$
145.9	134.2	49746.5	49485.1	104.7	48550.6

Do obliczenia związku między zmiennymi zastosujemy metodę korelacji liniowej.

Najpierw policzymy zależność między ekspresją CD62 i ekspresją PAC-1 dla płytek w stanie spoczynkowym. Skorzystamy z równania na współczynnik korelacji:

$$r = \frac{\left[\sum xy - \left(\sum x\right)\left(\sum y\right)/n\right]}{\sqrt{\left[\sum x^2 - \left(\sum x\right)^2/n\right] * \left[\sum y^2 - \left(\sum y\right)^2/n\right]}} =$$

$$= \frac{(104.7) - (40.9)(37.1)/15}{\sqrt{\left[145.9 - (40.9)^2/15\right] * \left[134.2 - (37.1)^2/15\right]}} = \frac{3.54067}{\sqrt{(34.3793)(42.4393)}} = 0.093$$

Aby dowiedzieć się, czy obliczony współczynnik korelacji jest istotny (to znaczy, czy występuje rzeczywista zależność między zmiennymi), postawione poniżej hipotezy postanowiono zweryfikować przy poziomie istotności $\alpha = 0.05$:

– H_0: obserwowana relacja między zmiennymi jest przypadkowa, czyli rzeczywiste $r_{xy} = 0$

– H_A: obserwowana relacja między zmiennymi nie jest przypadkowa, czyli rzeczywiste $r_{xy} \neq 0$

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.093\sqrt{15-2}}{\sqrt{1-(0.093)^2}} = \frac{0.3353}{0.99567} = 0.337$$

Ponieważ $0.337 = t_{dośw} < t_{0.05(2),13} = 2.160$, nie mamy podstaw do odrzucenia hipotezy zerowej: wnioskujemy, że nie ma związku między zmiennymi.

Następnie policzmy zależność między ekspresją CD62 oraz ekspresją PAC-1 dla płytek aktywowanych:

$$r = \frac{\left[\sum xy - \left(\sum x\right)\left(\sum y\right)/n\right]}{\sqrt{\left[\sum x^2 - \left(\sum x\right)^2/n\right] * \left[\sum y^2 - \left(\sum y\right)^2/n\right]}} =$$

$$= \frac{(48550.6) - (825.3)(832.1)/15}{\sqrt{\left[49746.5 - (825.3)^2/15\right] * \left[49485.1 - (832.1)^2/15\right]}} = \frac{2768.458}{\sqrt{(4338.494)(3325.7393)}} = 0.729$$

Policzmy, czy obliczony współczynnik korelacji jest istotny (czy występuje rzeczywista zależność między zmiennymi). Zweryfikujmy postawione poniżej hipotezy przy poziomie istotności $\alpha = 0.05$:

– H_0: obserwowana relacja między zmiennymi jest przypadkowa, czyli rzeczywiste $r_{xy} = 0$

– H_A: obserwowana relacja między zmiennymi nie jest przypadkowa, czyli rzeczywiste $r_{xy} \neq 0$

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.729\sqrt{15-2}}{\sqrt{1-(0.729)^2}} = \frac{2.628447}{0.46856} = 5.61$$

Ponieważ $5.61 = t_{dośw} > t_{0.05(2),13} = 2.160$, odrzucamy hipotezę zerową i wnioskujemy, że związek między zmiennymi nie jest przypadkowy.

Przykład 63

Porównywano dwie metody oznaczania fosfotyrozyny techniką Western immunoblotting: z peroksydazą chrzanową oraz fosfatazą alkaliczną. Czy wyniki uzyskiwane obiema metodami (*jedn. umowne*) wykazują korelację?

	metoda 1	**metoda 2**	x^2	y^2	xy	$\Delta_1 = x_i - \bar{x}$	$\Delta_2 = y_i - \bar{y}$	$\Delta_1 * \Delta_2$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
	52	97	2704	9409	5044	–5.7	–8.5	48.54	32.87	71.68
	36	78	1296	6084	2808	–21.7	–27.5	596.94	472.34	754.42
	44	84	1936	7056	3696	–13.7	–21.5	294.81	188.60	460.82
	55	112	3025	12544	6160	–2.7	6.5	–17.86	7.47	42.68
	53	102	2809	10404	5406	–4.7	–3.5	16.41	22.40	12.02
	67	112	4489	12544	7504	9.3	6.5	60.54	85.87	42.68
	72	130	5184	16900	9360	14.3	24.5	350.01	203.54	601.88
	55	90	3025	8100	4950	–2.7	–15.5	42.28	7.47	239.22
	66	117	4356	13689	7722	8.3	11.5	95.34	68.34	133.02
	46	94	2116	8836	4324	–11.7	–11.5	134.54	137.67	131.48
	77	124	5929	15376	9548	19.3	18.5	357.08	371.20	343.48
	57	105	3249	11025	5985	–0.7	–0.5	0.34	0.54	0.22
	59	115	3481	13225	6785	1.3	9.5	12.08	1.60	90.88
	70	125	4900	15625	8750	12.3	19.5	239.61	150.47	381.55
	57	97	3249	9409	5529	–0.7	–8.5	6.21	0.54	71.68
	$\bar{x} = 57.7$	$\bar{y} = 105.5$								
$\sum$	866	1582	51748	170226	93571	0	0	2236.87	1750.93	3377.73

Zastosujmy pierwszy sposób obliczeń:

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\left[\sum (x - \bar{x})^2 \sum (y - \bar{y})^2\right]}} = \frac{2236.87}{\sqrt{(1750.93)(3377.73)}} = \frac{2236.87}{2431.91} = 0.920$$

Jeżeli zastosujemy drugą formułę liczenia *r* (por. str. 102), postępujemy następująco:

$$r = \frac{\left[\sum xy - \left(\sum x\right)\left(\sum y\right)/n\right]}{\sqrt{\left[\sum x^2 - \left(\sum x\right)^2/n\right] * \left[\sum y^2 - \left(\sum y\right)^2/n\right]}} =$$

$$= \frac{(93571) - (866)(1582)/15}{\sqrt{\left[51748 - (866)^2/15\right] * \left[170226 - (1582)^2/15\right]}} = \frac{2236.867}{\sqrt{(1750.933)(3377.73)}} =$$

$$= \frac{2236.867}{2431.91} = 0.920$$

Postawione poniżej hipotezy postanowiono zweryfikować przy poziomie istotności $\alpha = 0.05$:

– H_0: obserwowana relacja między zmiennymi jest przypadkowa, czyli rzeczywiste $r_{xy} = 0$.
– H_A: obserwowana relacja między zmiennymi nie jest przypadkowa, czyli rzeczywiste $r_{xy} \neq 0$.

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.920\sqrt{15-2}}{\sqrt{1-(0.920)^2}} = \frac{3.317}{0.3919} = 8.4635$$

Ponieważ $8.4635 = t_{dośw} > t_{0.05(2),13} = 2.160$, odrzucamy hipotezę zerową i wnioskujemy, że związek między zmiennymi nie jest przypadkowy.

Przykład 64

W tabeli poniżej podano stopień zahamowania wzrostu kultur komórkowych w czasie. Czy te dwie zmienne są skorelowane?

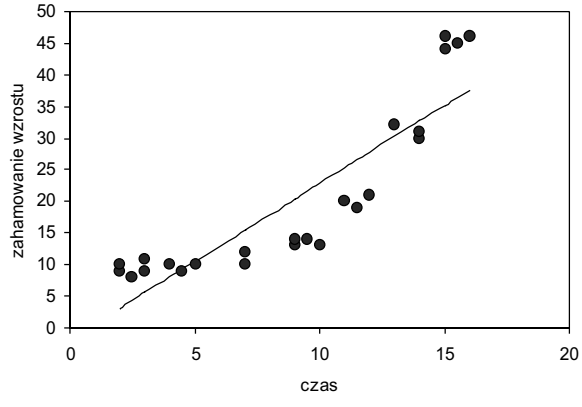
x_i (czas)	y_i	x_i^2	y_i^2	$x_i y_i$
2.0	9.00	4.00	81.0	18.0
2.0	10.00	4.00	100.0	20.0
2.5	8.00	6.25	64.0	20.0
3.0	9.00	9.00	81.0	27.0
3.0	11.00	9.00	121.0	33.0
4.0	10.00	16.00	100.0	40.0
4.5	9.00	20.25	81.0	40.5
5.0	10.00	25.00	100.0	50.0
7.0	10.00	49.00	100.0	70.0
7.0	12.00	49.00	144.0	84.0
9.0	13.00	81.00	169.0	117.0
9.0	14.00	81.00	196.0	126.0
9.5	14.00	90.25	196.0	133.0
10.0	13.00	100.00	169.0	130.0
11.0	20.00	121.00	400.0	220.0
11.5	19.00	132.25	361.0	218.5
12.0	21.00	144.00	441.0	252.0

	13.0	32.00	169.00	1024.0	416.0
	14.0	30.00	196.00	900.0	420.0
	14.0	31.00	196.00	961.0	434.0
	15.0	44.00	225.00	1936.0	660.0
	15.0	46.00	225.00	2116.0	690.0
	15.5	45.00	240.25	2025.0	697.5
	16.0	46.00	256.00	2116.0	736.0
średnia	8.9375	20.25			
$\sum =$	214.5	486.00	2448.25	13982.0	5652.5

Współczynnik korelacji liniowej Pearsona policzony dla takiej pary zmiennych wynosi:

$$r = \frac{\left[\sum xy - \left(\sum x\right)\left(\sum y\right)/n\right]}{\sqrt{\left[\sum x^2 - \left(\sum x\right)^2/n\right] * \left[\sum y^2 - \left(\sum y\right)^2/n\right]}} =$$

$$= \frac{[5652.5 - (214.5)(486)/24]}{\sqrt{[2448.25 - (214.5)^2/24] * [13982 - (486)^2/24]}} = \frac{1308.875}{1482.99} = 0.8826$$



Jeżeli jednak przyjrzymy się rozkładowi punktów w układzie współrzędnych, widzimy, że punkty te nie rozkładają się zupełnie na linii prostej, opisują one raczej krzywą wykładniczą; z uwagi na duże odchylenia punktów od linii prostej współczynnik korelacji liniowej jest silnie zaniżony. O wiele lepszym rozwiązaniem byłoby policzenie dla tego przypadku korelacji nieliniowej. Chcąc się posłużyć metodą oszacowania współczynnika eta korelacji nieliniowej musimy pogrupować dane w kategorie. Widzimy na wykresie, że punkty tworzą przynajmniej siedem takich zgrupowań. Każde takie zgrupowanie punktów doświadczalnych jest jakby krytycznym regionem „załamania się" liniowości rozkładu danych.

x_i (czas)	y_i	$\overline{y}_c$	$y_i - \overline{y}_c$	$(y_i - \overline{y}_c)^2$	$y_i - \overline{y}_t$	$(y_i - \overline{y}_t)^2$
2.0	9		–0.4000	0.16000	–11.25	126.5625
2.0	10		0.6000	0.36000	–10.25	105.0625
2.5	8		–1.4000	1.96000	–12.25	150.0625
3.0	9		–0.4000	0.16000	–11.25	126.5625
3.0	11	9.400	1.6000	2.56000	–9.25	85.5625
4.0	10		0.3333	0.11110	–10.25	105.0625
4.5	9		–0.6667	0.44440	–11.25	126.5625
5.0	10	9.667	0.3333	0.11110	–10.25	105.0625
7.0	10		–1.0000	1.00000	–10.25	105.0625
7.0	12	11.000	1.0000	1.00000	–8.25	68.0625
9.0	13		–0.5000	0.25000	–7.25	52.5625
9.0	14		0.5000	0.25000	–6.25	39.0625
9.5	14		0.5000	0.25000	–6.25	39.0625
10.0	13	13.500	–0.5000	0.25000	–7.25	52.5625
11.0	20		0.0000	0.00000	–0.25	0.0625
11.5	19		–1.0000	1.00000	–1.25	1.5625
12.0	21	20.000	1.0000	1.00000	0.75	0.5625
13.0	32		1.0000	1.00000	11.75	138.0625
14.0	30		–1.0000	1.00000	9.75	95.0625
14.0	31	31.000	0.0000	0.00000	10.75	115.5625
15.0	44		–1.2500	1.56250	23.75	564.0625
15.0	46		0.7500	0.56250	25.75	663.0625
15.5	45		–0.2500	0.06250	24.75	612.5625
16.0	46	45.250	0.7500	0.56250	25.75	663.0625
		$\sum$ =	0.0000	15.61667	0.00	4140.5000

Obliczamy współczynnik korelacji eta (η):

$$\eta = \sqrt{1 - \frac{\sum (y_i - \overline{y}_c)^2}{\sum (y_i - \overline{y}_t)^2}} = \sqrt{1 - \frac{15.61667}{4140.5}} = \sqrt{1 - 0.0037716} = 0.9981$$

Zauważmy, że jest on o wiele wyższy niż współczynnik korelacji liniowej (*r*).

Przykład 65

W tabeli podano miary aktywacji płytek krwi w zakresie trzech markerów aktywacji mierzonych metodą cytometrii przepływowej:

(*a*) selektyna P	(*b*) aktywny GPIIbIIIa	(*c*) GPIb
43.5	23.5	77.2
35.7	18.2	82.6
39.2	20.6	79.5
61.7	35.7	59.4
55.2	31.2	65.1
48.4	27.1	71.9
53.9	27.9	63.8
66.4	33.8	63.2
50.6	28.4	61.5
44.3	22.5	74.5

a) Jaka jest rzeczywista korelacja między zmianami ekspresji selektyny P i GPIb po usunięciu efektu zmian ekspresji aktywnego kompleksu GPIIbIIIa?
b) Jaka jest rzeczywista korelacja między zmianami ekspresji aktywnego kompleksu GPIIbIIIa i GPIb po usunięciu efektu zmian ekspresji selektyny P?

Aby odpowiedzieć na te pytania, powinniśmy zastosować metodę korelacji cząstkowej. Policzmy najpierw współczynniki zwykłej korelacji liniowej między parametrami. Są one następujące:

	selektyna P	**aktywny GPIIbIIIa**	**GPIb**
selektyna P		0.969	–0.897
aktywny GPIIbIIIa	0.969		–0.925
GPIb	–0.897	–0.925	

Pytanie a)

Współczynnik korelacji cząstkowej wynosi:

$$r_{ac,b} = \frac{r_{ac} - (r_{ab})(r_{bc})}{\sqrt{(1 - r_{ab}^2)(1 - r_{bc}^2)}}$$

$$r_{ac,b} = \frac{(-0.897) - (0.969)(-0.925)}{\sqrt{(1 - (0.969)^2)(1 - (-0.925)^2)}} = \frac{-0.001}{0.094495} = -0.01058$$

Tyle wynosi korelacja między zmianami ekspresji selektyny P i GPIb po usunięciu efektu zmian ekspresji aktywnego kompleksu GPIIbIIIa, czyli w sytuacji, gdyby ekspresja aktywnego kompleksu GPIIbIIIa pozostawała stała (nie zmieniała się).

Pytanie b)

$$r_{bc,a} = \frac{r_{bc} - (r_{ab})(r_{ac})}{\sqrt{(1 - r_{ab}^2)(1 - r_{ac}^2)}}$$

$$r_{bc,a} = \frac{(-0.925) - (0.969)(-0.897)}{\sqrt{(1-(0.969)^2)(1-(-0.897)^2)}} = \frac{-0.05615}{0.109916} = -0.511$$

Tyle wynosiłaby korelacja między zmianami ekspresji aktywnego kompleksu GPIIbIIIa i GPIb, gdyby ekspresja selektyny P pozostawała stała.

Wartość statystyki testu *t* dla każdej z tych korelacji cząstkowych wynosi:

$$t_{ac,b} = \frac{r_{ac,b}\sqrt{n-k-1}}{\sqrt{1-(r_{ac,b})^2}} = \frac{(-0.01058)\sqrt{10-2-1}}{\sqrt{1-(-0.01058)^2}} = \frac{-0.028}{0.999944} = -0.028$$

Ponieważ $t_{0.05(1),8} = 2.306 > t_{dośw}$, nie mamy podstaw do odrzucenia hipotezy zerowej mówiącej, że korelacja cząstkowa między zmiennymi *a* i *c* (selektyna P i GPIb) jest nieistotna. Zatem nie występuje rzeczywisty związek między ekspresją selektyny P i ekspresją GPIb. Co to oznacza w świetle wyniku istotnej i wysokiej korelacji liniowej między tymi zmiennymi? Istotna korelacja liniowa Pearsona jakiejś zmiennej i nieistotna korelacja cząstkową tej zmiennej odzwierciedla sytuację, gdy wpływ danej zmiennej jest w pełni tłumaczony przez inne zmienne. W naszym przypadku – gdybyśmy uwzględnili zależności pozostałych parametrów (np. GPIIIa), wtedy ta rozważana korelacja (selektyna P *vs*. GPIb) przestaje być istotna i nie musimy jej uwzględniać w modelu:

$$t_{bc,a} = \frac{r_{bc,a}\sqrt{n-k-1}}{\sqrt{1-(r_{bc,a})^2}} = \frac{(-0.511)\sqrt{10-2-1}}{\sqrt{1-(-0.511)^2}} = \frac{-1.35167}{0.859651} = -1.57234$$

Ponieważ $t_{0.05(1),8} = 2.306 > t_{dośw}$, nie mamy podstaw do odrzucenia hipotezy zerowej mówiącej, że korelacja cząstkowa między zmiennymi *b* i *c* (GPIIIa i GPIb) jest nieistotna. Możemy zatem uznać, że nie występuje rzeczywisty związek między ekspresją glikoprotein IIIa i Ib, a tym samym – że związek ten nie tłumaczy także występowania liniowej korelacji między ekspresją selektyny P i ekspresją GPIb.

Regresja liniowa i wielokrotna

Przykład 66

Poniżej przedstawiono wyniki absorbancji dla różnych stężeń roztworu barwnej substancji.

	stężenie	absorbancja	x^2	y^2	xy
	1	0.24	1	0.0576	0.24
	2	0.66	4	0.4356	1.32
	4	1.15	16	1.3225	4.60
	8	2.34	64	5.4756	18.72
$\sum =$	15	4.39	85	7.2913	24.88

Czy istnieje relacja między stężeniem a wartością absorbancji roztworu? Jeżeli tak, to należy przedstawić wykres tej zależności z zaznaczeniem 95% przedziału ufności.

Postawmy poniższe hipotezy, które zweryfikujemy przy poziomie istotności $\alpha = 0.05$.
– H_0: obserwowana relacja między zmiennymi jest przypadkowa, czyli rzeczywiste $r_{xy} = 0$
– H_A: obserwowana relacja między zmiennymi nie jest przypadkowa, czyli rzeczywiste $r_{xy} \neq 0$
Hipotezę zerową odrzucimy, jeżeli $F_{dośw} > F_{0.05(1),2} = 18.5$.
Obliczamy współczynniki równania:

$$b = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - \left(\sum x\right)^2} \qquad \text{oraz} \qquad a = \frac{\sum y - b\sum x}{n}$$

$$b = \frac{4(24.88) - (15)(4.39)}{4(85) - (15)^2} = 0.292783 \qquad a = \frac{4.39 - 0.292783\,(15)}{4} = -0.0004347$$

Obliczamy współczynnik determinacji:

$$SS_{całk} = \sum y^2 - \frac{\left(\sum x\right)^2}{n} = 7.2913 - \frac{(4.39)^2}{4} = 7.2913 - \frac{19.2721}{4} = 2.473275$$

$$SS_{wyjaśniona} = b^2 * \left[\sum x^2 - \frac{\left(\sum x\right)^2}{n}\right] = (0.292783)^2 * \left[85 - \frac{(15)^2}{4}\right] = 2.4645042$$

$$SS_{niewyjaśniona} = SS_{całk} - SS_{wyjaśniona} = 2.473275 - 2.4645042 = 0.0087708$$

$$r^2 = \frac{SS_{wyjaśniona}}{SS_{całkowita}} = \frac{2.4645042}{2.473275} = 0.99645$$

Zestawiamy tabele analizy wariancji:

zmienność	*df*	*SS*	*MS*	*F*	**istotność *F***
regresja	1	2.464498	2.464498	561.556	0.0018
resztkowy	2	0.008777	0.004389		
całkowita	3	2.473275			

Ponieważ $561.556 = F_{dośw} > F_{0.05(1),2} = 18.5$, hipotezę zerową odrzucamy i wnioskujemy, że istnieje zależność liniowa między stężeniem a wartością absorbancji.
Testujemy istotność współczynnika kierunkowego prostej regresji, β:
H_0: $\beta = 0$,
H_A: $\beta \neq 0$.

Hipotezę zerową odrzucimy, jeżeli $t_{dośw} > t_{0.05(1),2} = 4.302$.

$$s = \sqrt{\left[\frac{\sum\left(y-\bar{y}\right)^2 - b^2\sum\left(x-\bar{x}\right)^2}{(n-2)}\right]} =$$

$$= \sqrt{\frac{\left[\sum y^2 - \left(\sum y\right)^2/n\right] - b^2 * \left[\sum x^2 - \left(\sum x\right)^2/n\right]}{(n-2)}} =$$

$$= \sqrt{\frac{[7.2913 - (4.39)^2/4] - (0.292783)^2 * [85 - (15)^2/4]}{2}} = 0.0662223$$

$$SE_b = \frac{s}{\sqrt{\sum\left(x-\bar{x}\right)^2}} =$$

$$= \frac{s}{\sqrt{\sum x^2 - \left(\sum x\right)^2/n}} = \frac{0.0662223}{\sqrt{85-(15)^2/4}} = \frac{0.0662223}{5.3619} = 0.0123505$$

$$t = \frac{b-\beta}{SE_b} = \frac{0.292783 - 0}{0.0123505} = 23.706 \qquad \text{przy} \quad d.f. = n-2 = 2 \text{ stopniach swobody}$$

Ponieważ $23.706 = t_{dośw} > t_{0.05(1),2} = 4.302$, hipotezę zerową odrzucamy i wnioskujemy, że nachylenie prostej jest istotne oraz istnieje zależność liniowa między stężeniem a wartością absorbancji.

Obliczamy 95% przedział ufności współczynnika β:

$$\beta = b \pm t_{0.05(1),2} \text{ x } SE_b = 0.292783 \pm 4.302(0.0123505) = 0.292783 \pm 0.053132$$

$$0.240 < \beta < 0.346$$

Z prawdopodobieństwem 95% możemy powiedzieć, że nachylenie prostej leży w przedziale od 0.240 do 0.346.

Obliczamy granice przedziału ufności dla prostej regresji:

$$\bar{y} = y_c \pm t_{\alpha(1),n-2} * \sqrt{MS_{resz}} * \sqrt{\frac{1}{n} + \frac{\left(x_i-\bar{x}\right)^2}{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}}$$

Na przykład dla stężenia $x = 1$:

$$\bar{y} = 0.292 \pm 4.302\sqrt{0.004389}\sqrt{\frac{1}{4} + \frac{(1-3.75)^2}{85 - \frac{(15)^2}{4}}}$$

$$\bar{y} = 0.292 \pm 0.204 \qquad 0.088 < \bar{y} < 0.496$$

Podobnie, dla stężenia $x = 4$:

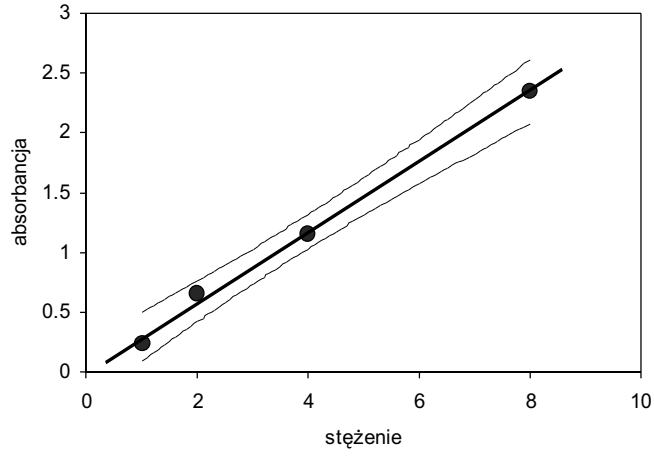
$$\bar{y} = 1.171 \pm 4.302\sqrt{0.004389}\sqrt{\frac{1}{4} + \frac{(4-3.75)^2}{85 - \frac{(15)^2}{4}}}$$

$$\bar{y} = 1.171 \pm 0.143 \quad 1.028 < \bar{y} < 1.314$$

Wyniki dla 95% przedziału ufności:

stężenie	absorbancja	xy	y_c	dolny limit	górny limit	zakres
1	0.24	0.24	0.292	0.088	0.496	0.408
2	0.66	1.32	0.585	0.415	0.755	0.340
4	1.15	4.6	1.171	1.028	1.314	0.286
8	2.34	18.72	2.342	2.075	2.609	0.534

Zauważmy, że granice oraz zakres przedziału ufności dla prostej regresji nie jest jednakowy dla różnych wartości zmiennej x; wśród przedstawionych danych jest on najmniejszy dla $x = 4$. Oznacza to, że dokładność dopasowania krzywej teoretycznej do punktów doświadczalnych jest największa w tym regionie prostej regresji.



Przykład 67

Badano zależność stężenia całkowitego cholesterolu w osoczu krwi u pacjentów z chorobą wieńcową od stężenia cholesterolu we frakcjach lipoprotein LDL i HDL[1]. Aby zaobserwować różnice między metodą regresji prostej i regresji wielorakiej przeanalizujmy poniższe opracowanie danych. Policzymy najpierw parametry regresji oraz parametry analizy wariancji przy założeniu badania wpływu każdej ze zmiennych (LDL, HDL) osobno (tzn. w modelu regresji prostej). Przejrzyjmy i porównajmy tabele poniżej.

Podsumowanie regresji zmiennej zależnej: TC vs. LDL[2]

$r = 0.939$ $r^2 = 0.881$
$F_{1,207} = 1530.2$ $p« 0.00001$ błąd std. estymacji: 16.312

	wartość	**błąd std.**	$t_{(207)}$	**istotność** (p)
a	75.845	4.198	18.07	0.000
b_1 (LDL)	0.999	0.026	39.12	0.000

Analiza wariancji: TC vs. LDL

zmienność	***SS***	***d.f.***	***MS***	***F***	**istotność**
regresja	407134.4	1	407134.4	1530.202	0.00
resztowa	55075.6	207	266.1		
całkowita	462210.0				

Podsumowanie regresji zmiennej zależnej: TC vs. HDL

$r = 0.076$ $r^2 = 0.0058$
$F_{1,207} = 1.205$ $p< 0.274$ błąd std. estymacji: 47.12

	wartość	**błąd std.**	$t_{(207)}$	**istotność** (p)
a	221.55	11.803	18.77	0.000
b_2 (HDL)	0.276	0.252	1.098	0.274

Analiza wariancji: TC vs. HDL

zmienność	***SS***	***d.f.***	***MS***	***F***	**istotność**
regresja	2675.0	1	2675.030	1.204981	0.273604
resztowa	459535.0	207	2219.976		
całkowita	462210.0				

Widzimy, że zmienność całkowita pozostaje stała dla każdej ze zmiennych niezależnych, różne są natomiast udziały zmienności wyjaśnionej (przez regresję) oraz zmienności niewyjaśnionej (resztowej). Dla naszego przypadku największą zmienność wyjaśnioną zanotowaliśmy dla LDL; oznacza to, że ta zmienna w największym stopniu wpływa na

[1] Dane dla tego przykładu znajdzie Czytelnik w zbiorach ***regr-wielokrotna.xls*** oraz ***regr-wielokrotna.sta***.
[2] Obliczenia wykonano przy użyciu pakietu statystycznego Statistica PL ver. 5.3 (Statsoft Polska).

dopasowanie modelu. Ta zmienna przyczynia się także w największym stopniu do wiarygodnej predykcji (przewidywania) badanej zmiennej zależnej (TC), gdyż posiada najwyższą wartość współczynnika kierunkowego *b*.

Jeżeli włączymy do modelu obie zmienne niezależne razem to uzyskamy następujące wyniki:

Podsumowanie regresji zmiennej zależnej: TC vs. LDL i HDL

$r = 0.956$ $r^2 = 0.914$
$F_{2,206} = 1098.0$ $p \ll 0.00001$ błąd std. estymacji: 13.872

	wartość	**błąd std.**	$t_{(207)}$	**istotność** (*p*)
a	42.257	5.178	8.162	0.000
b_1 (LDL)	1.02095	0.0219	46.713	0.000
b_2 (HDL)	0.66813	0.0746	8.956	0.000

Analiza wariancji: TC vs. LDL i HDL

zmienność	***SS***	***d.f.***	***MS***	***F***	**istotność**
regresja	422570.2	2	211285.1	1098.006	0.000
resztowa	39639.8	206	192.4		
całkowita	462210.0				

W modelu tym zmienność wyjaśniona (przyporządkowana regresji) odpowiada za ponad 91% (422570.2/462210) zmienności całkowitej. Proporcja ta jest liczbowo równa wartości współczynnika determinacji r^2, gdzie $r = \sqrt{0.9142385} = 0.95616$ nazywany jest współczynnikiem korelacji wielokrotnej. Wartość współczynnika korelacji wielokrotnej jest zawsze dodatnia, gdyż nie można przypisać określonego kierunku (prostej lub odwrotnej proporcjonalności) wspólnej (wzajemnej) korelacji (zależności) więcej niż dwóch zmiennych.

Na całkowitą zmienność regresji dwóch zmiennych niezależnych (LDL i HDL) oraz zmiennej zależnej TC składa się zmienność dotycząca LDL, jak również pewna dodatkowa porcja zmienności wyjaśnianej przez zmienność parametru LDL skorygowanego na obecność parametru HDL. Istotność tej dodatkowej porcji zmienności w przypadku LDL ($F_{1,206} = 80.22$, $p \ll 0.0001$) mówi nam o tym, że model pozostaje dopasowany także w obecności obu zmiennych, chociaż dopasowanie to nie polepsza się w stosunku do tego, jakie obserwujemy w obecności jedynie zmiennej LDL. Analogicznie, jeżeli podobnie przeanalizujemy wpływ takiej dodatkowej porcji zmienności wyjaśnianej przez zmienność HDL skorygowaną na obecność LDL, to widzimy, że obecność dwóch zmiennych polepsza dopasowanie modelu, czyli możemy powiedzieć, że włączenie zmiennej LDL i zmiennej HDL wpływa na lepsze dopasowanie modelu w stosunku do sytuacji, gdy obecna jest w modelu jedynie zmienna HDL.

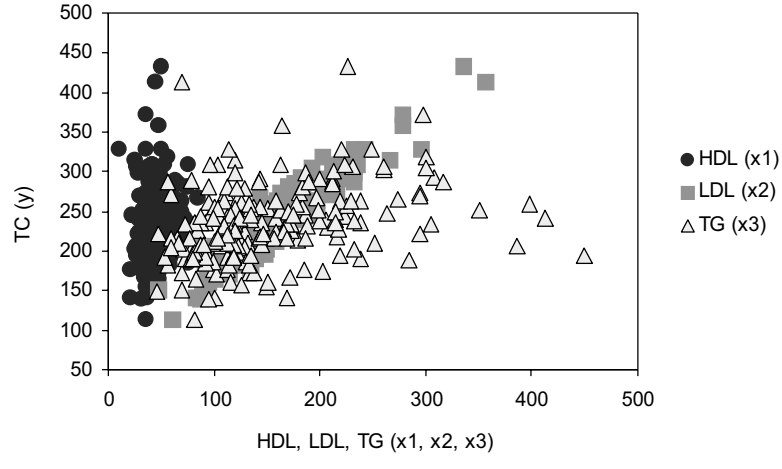
zmienność regresji	*SS*	*d.f.*	*MS*	*F*	**istotność**
całkowita	422570.2	2			
LDL	407134.4	1	407134.4	2115.79	0.00010
LDL skorygowana na obecność HDL	15435.8	1	15435.8	80.22	0.00010
HDL	2675.0	1	2675.0	13.90	0.00025
HDL skorygowana na obecność LDL	419895.2	1	419895.2	2182.11	0.00010
resztowa	39639.8	206	192.40		

Ponieważ dwie zmienne: LDL i HDL są wzajemnie skorelowane ze sobą, zmienność w modelu regresji wielokrotnej (tzn. przy uwzględnieniu jednoczesnego wpływu kilku zmiennych niezależnych/wyjaśniających) na zmienną zależną nie jest prostą sumą zmienności każdego z parametrów ocenianą w modelu regresji prostej (tzn. 407134.4 + 2675.0 ≠ 422570.2).

Porównywanie współczynników regresji

Przykład 68

Rozkłady danych doświadczalnych (dane w arkuszu ***regr-wielokrotna.xls***) stężeń całkowitego cholesterolu w osoczu w funkcji stężenia cholesterolu frakcji HDL (x_1), LDL (x_2) i TG (x_3) przedstawia rysunek:

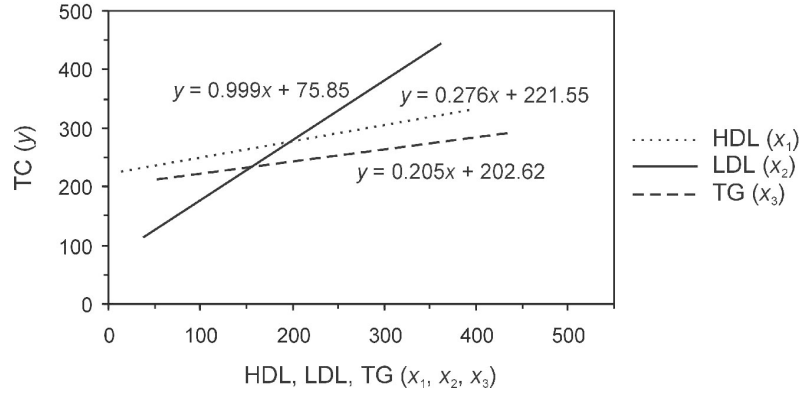


Należy ocenić równoległość prostych opisujących zależności TC = $a_2 + b_2$x (LDL) względem TC = $a_3 + b_3$(TG) oraz TC = $a_1 + b_1$x (HDL) względem TC = $a_3 + b_3$(TG).

Obliczamy współczynniki równań regresji opisujących zależności każdej ze zmiennych niezależnych ze zmienną zależną.

	$\sum x$	$\sum y$	$\sum x^2$	$\sum xy$	$\sum y^2$	N	a	b
TC		48906			11906214	209		
HDL	9417.0		459327	2213257		209	221.55	0.276
LDL	33089.4		5646794	8150486		209	75.85	0.999
TG	32011.0		5999809	7715282		209	202.62	0.205

Proste regresji opisujące zależności przedstawia rysunek:



Testujemy pary hipotez postaci:
H_0: $b_1 = b_2$
H_A: $b_1 \neq b_2$

Obliczenia:
Dla zmiennej LDL:

$$\sum x^2 = 5646794 \qquad \sum xy = 8150486 \qquad \sum y^2 = 11906214$$

$$n = 209, \qquad b = 0.999$$

$$SS_{reszt} = \sum y^2 - \frac{(\sum xy)^2}{\sum x^2} = 11906214 - \frac{(8150486)^2}{5646794} = 14194.74$$

$$df_{reszt} = 209 - 2 = 207$$

Dla zmiennej TG:

$$\sum x^2 = 5999809 \qquad \sum xy = 7715282 \qquad \sum y^2 = 11906214$$

$$n = 209, \qquad b = 0.205$$

$$SS_{reszt} = \sum y^2 - \frac{(\sum xy)^2}{\sum x^2} = 11906214 - \frac{(7715282)^2}{5999809} = 1984968.8$$

$$df_{reszt} = 209 - 2 = 207$$

Dla zmiennej HDL:

$$\sum x^2 = 459327 \qquad \sum xy = 2213257 \qquad \sum y^2 = 11906214$$

$$n = 209, \qquad b = 0.276$$

$$SS_{reszt} = \sum y^2 - \frac{(\sum xy)^2}{\sum x^2} = 11906214 - \frac{(2213257)^2}{459327} = 10664530$$

$$df_{reszt} = 209 - 2 = 207$$

TC *vs.* LDL względem TC *vs.* TG:

$$\left(s^2_{x,y}\right)_p = \frac{(SS_{reszt})_1 + (SS_{reszt})_2}{(df_{reszt})_1 + (df_{reszt})_2} = \frac{141941.74 + 1984968.8}{207 + 207} = \frac{2126910.5}{414} = 5137.465$$

$$SE_{b_1 - b_2} = \sqrt{\frac{\left(s^2_{x,y}\right)_p}{\sum x_1^2} + \frac{\left(s^2_{x,y}\right)_p}{\sum x_2^2}} =$$

$$= \sqrt{\frac{5137.465}{5646794} + \frac{5137.465}{5999809}} = \sqrt{0.0009098 + 0.0008562} = 0.0420246$$

$$t = \frac{b_1 - b_2}{SE_{b_1 - b_2}} = \frac{0.999 - 0.205}{0.0420246} = 18.89 \qquad v = 207 + 207 = 414$$

$t_{dośw} > t_{0.0001(2),414}$, zatem odrzucamy hipotezę zerową i wnioskujemy, że dwie porównywane linie regresji nie są równoległe. Skoro linie nie są równoległe to możemy obliczyć punkt ich przecięcia jako:

$$x_{przecięcia} = \frac{a_2 - a_1}{b_1 - b_2} \quad \text{oraz} \quad y_{przecięcia} = a_1 + b_1 x_1 \quad \text{lub} \quad y_{przecięcia} = a_2 + b_2 x_1$$

$$x_{przecięcia} = \frac{a_2 - a_1}{b_1 - b_2} = \frac{202.62 - 75.85}{0.999 - 0.205} = \frac{126.77}{0.794} = 159.66$$

$$y_{przecięcia} = 202.62 + 0.205 * 159.66 = 235.35$$

Współrzędne punktu przecięcia dwóch linii regresji wynoszą (*x*,y) = (159.66, 235.35).

TC *vs*. HDL względem TC *vs*. TG:

$$\left(s^2_{x,y}\right)_p = \frac{(SS_{reszt})_1 + (SS_{reszt})_2}{(df_{reszt})_1 + (df_{reszt})_2} = \frac{10664530 + 1984968.8}{207 + 207} = \frac{12649499}{414} = 30554.345$$

$$SE_{b_1 - b_2} = \sqrt{\frac{\left(s^2_{x,y}\right)_p}{\sum x_1^2} + \frac{\left(s^2_{x,y}\right)_p}{\sum x_2^2}} =$$

$$= \sqrt{\frac{30554.345}{459327} + \frac{30554.345}{5999809}} = \sqrt{0.0716123} = 0.2676056$$

$$t = \frac{b_1 - b_2}{SE_{b_1 - b_2}} = \frac{0.276 - 0.205}{0.0420246} = 0.2653 \qquad v = 207 + 207 = 414$$

$t_{dośw} < t_{0.05(2),414}$, zatem nie mamy podstaw do odrzucenia hipotezy zerowej i wnioskujemy, że dwie porównywane linie regresji są równoległe.

Przykład 69

Korzystając z tych samych danych należy zbadać istotności różnic współczynników a_i dla par linii regresji TC *vs*. LDL i TC *vs*. HDL oraz TC *vs*. HDL i TC *vs*. TG.

Dla zmiennej LDL mamy:

$$\sum x = 33089.4 \qquad \sum y = 48906$$

$$x = 158.32 \qquad y = 234$$

$$\sum x^2 = 5646794 \qquad \sum xy = 8150486 \qquad \sum y^2 = 11906214$$

$$n = 209 \quad b = 0.999 \quad a = 75.85$$

$$SS_{reszt} = \sum y^2 - \frac{(\sum xy)^2}{\sum x^2} = 11906214 - \frac{(8150486)^2}{5646794} = 141941.74$$

$$df_{reszt} = 209 - 2 = 207$$

Dla zmiennej TG mamy:

$$\sum x = 32011 \qquad \sum y = 48906$$

$$\bar{x} = 153.16 \qquad \bar{y} = 234$$

$$\sum x^2 = 5999809 \qquad \sum xy = 7715282 \qquad \sum y^2 = 11906214$$

$$n = 209 \qquad b = 0.205 \qquad a = 202.62$$

$$SS_{reszt} = \sum y^2 - \frac{(\sum xy)^2}{\sum x^2} = 11906214 - \frac{(7715282)^2}{5999809} = 1984968.8$$

$$df_{reszt} = 209 - 2 = 207$$

Dla zmiennej HDL:

$$\sum x = 9417 \qquad \sum y = 48906$$

$$\bar{x} = 45.06 \qquad \bar{y} = 234$$

$$\sum x^2 = 459327 \qquad \sum xy = 2213257 \qquad \sum y^2 = 11906214$$

$$n = 209 \qquad b = 0.276 \qquad a = 221.55 \qquad df_{reszt} = 209 - 2 = 207$$

Testujemy pary hipotez postaci:
H_0: $a_1 = a_2$
H_A: $a_1 \neq a_2$
Porównanie współczynników *a* linii regresji TC *vs*. LDL oraz TC *vs*. HDL:

$$A_c = \left(\sum x^2\right)_1 + \left(\sum x^2\right)_2 = 5646794 + 459327 = 6106121$$

$$B_c = \left(\sum xy\right)_1 + \left(\sum xy\right)_2 = 8150486 + 2213257 = 10363743$$

$$C_c = \left(\sum y^2\right)_1 + \left(\sum y^2\right)_2 = 11906214 + 11906214 = 23812428$$

$$SS_c = C_c - \frac{B_c^2}{A_c} = 23812428 - 17590082 = 6222346$$

$$df_c = n_1 + n_2 - 3 = 209 + 209 - 3 = 415$$

$$\left(s^2_{x,y}\right)_c = \frac{SS_c}{df_c} = 14993.6$$

Wspólny (ważony) współczynnik kierunkowy regresji liczymy z równania:

$$b_c = \frac{\left(\sum xy\right)_1 + \left(\sum xy\right)_2}{\left(\sum x^2\right)_1 + \left(\sum x^2\right)_2} \qquad b_c = \frac{8150486 + 2213257}{5646794 + 459327} = \frac{10363743}{6106121} = 1.697271$$

zaś wartość statystyki testu *t* dla porównania współczynników *a* według wzoru:

$$t = \frac{\left(\bar{y}_1 - \bar{y}_2\right) - b_c(\bar{x}_1 - \bar{x}_2)}{\sqrt{\left(s^2_{x,y}\right)_c \left[\frac{1}{n_1} + \frac{1}{n_2} + \frac{\left(\bar{x}_1 - \bar{x}_2\right)^2}{A_c}\right]}}$$

Analogicznie możemy przeprowadzić porównanie współczynników *a* dla równań regresji TC *vs*. HDL oraz TC *vs*. TG.
Podsumowanie wyników cząstkowych obliczeń przedstawiono w tabeli:

	(1) TC *vs*. HDL oraz TC *vs*. TG	**(2) TC *vs*. HDL oraz TC *vs*. LDL**
$\left(\sum xy\right)_1$	2213257	8150486
$\left(\sum xy\right)_2$	7715282	2213257
$\left(\sum x^2\right)_1$	459327	5646794
$\left(\sum x^2\right)_2$	5999809	459327
$\left(\sum y^2\right)_1$	11906214	11906214
$\left(\sum y^2\right)_2$	11906214	11906214
$\bar{y}_1$	234	234
$\bar{y}_2$	234	234
$\bar{x}_1$	45.06	158.32
$\bar{x}_2$	153.16	45.06
b_c	1.537131	1.697271
C_c	23812428	23812428
B_c	9928539	10363743
A_c	6459136	6106121
SS_c	8550962	6222346
df_c	415	415
$\left(s^2_{x,y}\right)_c$	20604.73	14993.6

$1/n_1$	0.004785	0.004785
$1/n_2$	0.004785	0.004785
$(\bar{x}_1 - \bar{x}_2)^2$	11685.61	12827.83
$U = \frac{(\bar{x}_1 - \bar{x}_2)^2}{A_c}$	0.001809	0.002101
$V = 1/n_1 + 1/n_2 + U$	0.011379	0.01167
$R = (s^2_{x,y})_c \times V$	234.4517	174.9783
$mianownik = \sqrt{R}$	15.31181	13.22793
$1/mianownik$	0.065309	0.075598
$(\bar{x}_1 - \bar{x}_2)$	–108.1	113.26
$b_c(\bar{x}_1 - \bar{x}_2)$	–166.164	192.2329
licznik	166.1639	–192.233
t	***10.852***	***–14.5323***

Ponieważ $|t_{dośw}| = 14.53 > t_{0.05(2),415} = 1.965$, możemy odrzucić hipotezę zerową mówiącą, że współczynniki *a* równań regresji TC *vs*. LDL oraz TC *vs*. HDL są równe.

Analogicznie, ponieważ $t_{dośw} = 10.85 > t_{0.05(2),415} = 1.965$, zatem możemy odrzucić hipotezę zerową mówiącą, że współczynniki *a* równań regresji TC *vs*. TG oraz TC *vs*. HDL są równe.

Przykład 70

Dla tych samych danych sprawdźmy, czy istnieje istotna różnica dla współrzędnych punktów leżących na prostych regresji:
TC = $a_1 + b_1$ x (LDL)
TC = $a_2 + b_2$ x (HDL)
przy wartości zmiennej niezależnej x = 155 mg/100 ml.
Testujemy parę hipotez:
H_0: $\mu_{y_1} = \mu_{y_2}$
H_A: $\mu_{y_1} \neq \mu_{y_2}$

Statystykę testu t porównania dwóch punktów leżących na dwóch liniach regresji obliczamy z równania:

$$t = \frac{y_1 - y_2}{s_{y_1 - y_2}}, \qquad \text{gdzie } s_{y_1-y_2} = \sqrt{(s^2_{x,y})_p \left[\frac{1}{n_1} + \frac{1}{n_2} + \frac{(x-\bar{x}_1)^2}{(\sum x^2)_1} + \frac{(x-\bar{x}_2)^2}{(\sum x^2)_2}\right]}$$

Korzystając z obliczeń w poprzednich przykładach możemy zapisać:
dla LDL:

$\bar{x}_1 = 158.32$ $\quad \sum x_1^2 = 5646794$ $\quad \sum x_1 y_1 = 8150486$ $\quad \sum y_1^2 = 11906214$

$n_1 = 209$ $\quad a_1 = 75.845$ $\quad b_1 = 0.999$

$$(SS_{reszt})_1 = \sum y_1^2 - \frac{\left(\sum x_1 y_1\right)^2}{\sum x_1^2} = 11906214 - 11764272 = 141941.74 \qquad df_{reszt} = 209 - 2 = 207$$

Wartość zmiennej zależnej w punkcie $x = 155$ wynosi:

$$y_1 = 75.845 + 0.999(155) = 230.68103$$

oraz dla HDL:

$$\bar{x}_2 = 45.06 \qquad \sum x_2^2 = 459327 \qquad \sum x_2 y_2 = 2213257 \qquad \sum y_2^2 = 11906214$$

$$n_2 = 209 \qquad a_2 = 221.55 \qquad b_2 = 0.276$$

$$(SS_{reszt})_2 = \sum y_2^2 - \frac{\left(\sum x_2 y_2\right)^2}{\sum x_2^2} = 11906214 - 10664530 = 1241684.1 \qquad df_{reszt} = 209 - 2 = 207$$

$$\left(s_{x,y}^2\right)_p = \frac{(SS_{reszt})_1 + (SS_{reszt})_2}{(df_{reszt})_1 + (df_{reszt})_2} = \frac{141941.74 + 12416841}{207 + 207} = \frac{13836258}{414} = 3342.0914$$

Wartość zmiennej zależnej w punkcie $x = 155$ wynosi:

$$y_2 = 221.55 + 0.276(155) = 264.38534$$

$$s_{y_1 - y_2} = \sqrt{3342.0914\left[\frac{1}{209} + \frac{1}{209} + \frac{(155 - 158.32)^2}{5646794} + \frac{(155 - 45.06)^2}{459327}\right]} = \sqrt{119.93} = 10.95$$

$$t = \frac{y_1 - y_2}{s_{y_1 - y_2}} = \frac{230.68 - 264.385}{10.95} = -3.08$$

liczba stopni swobody $\nu = 209 + 209 - 4 = 414$

Ponieważ $|t_{\text{dośw}}| = 3.08 > t_{0.005(2),415} = 2.823$, zatem możemy odrzucić hipotezę zerową i stwierdzić, że dwa punkty x, y_1 = (155; 230.7) oraz x, y_2 = (155; 264.39) leżące na dwóch liniach regresji mają istotnie różne współrzędne.

Przykład 71

Badano zależność hamowania uwalniania PAI-1 w hodowli komórek śródbłonka pod wpływem trzech różnych inhibitorów reduktazy hydroksymetyloglutarylo-CoA (HMG-CoA). Należy zbadać, czy zależność stężeniowa hamowania wzrostu komórek jest jednako-

wa dla różnych inhibitorów, to znaczy określić istotności różnic między współczynnikami *a* i *b* w trzech równaniach regresji:

inhibitor 1 $y_1 = a_1 + b_1 * x_1$
inhibitor 2 $y_2 = a_2 + b_2 * x_2$
inhibitor 3 $y_3 = a_3 + b_3 * x_3$

	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
inhibitor 1	1	2	*1*	*4*	*2*
	2	3	*4*	*9*	*6*
	3	5	*9*	*25*	*15*
	4	7	*16*	*49*	*28*
	5	9	*25*	*81*	*45*
	6	12	*36*	*144*	*72*
	7	16	*49*	*256*	*112*
	8	18	*64*	*324*	*144*
	9	25	*81*	*625*	*225*
	10	32	100	1024	320
inhibitor 2	10	2	*100*	*4*	*20*
	12	3	*144*	*9*	*36*
	14	6	*196*	*36*	*84*
	16	9	*256*	*81*	*144*
	18	11	*324*	*121*	*198*
	20	14	*400*	*196*	*280*
	22	19	*484*	*361*	*418*
	24	23	*576*	*529*	*552*
inhibitor 3	10	14	*100*	*196*	*140*
	15	18	*225*	*324*	*270*
	20	21	*400*	*441*	*420*
	25	31	*625*	*961*	*775*
	30	42	*900*	*1764*	*1260*
	35	51	*1225*	*2601*	*1785*
	40	60	*1600*	*3600*	*2400*
	45	66	*2025*	*4356*	*2970*
	50	78	*2500*	*6084*	*3900*
	$\sum x_i$	$\sum y_i$	$\sum x_i^2$	$\sum y_i^2$	$\sum x_i y_i$
inhibitor 1	*55*	*129*	*385*	*2541*	*969*
inhibitor 2	*136*	*87*	*2480*	*1337*	*1732*
inhibitor 3	*270*	*381*	*9600*	*20327*	*13920*
wspólna	**461**	**597**	**12465**	**24205**	**16621**
całkowita	*421*	*698*	*11865*	*29390*	*18421*

Wartości oznaczone kursywą zostały obliczone na podstawie danych wyjściowych i posłużyły do obliczenia innych, zaznaczonych pogrubionym drukiem.

Obliczamy parametry równań regresji:

Dla inhibitora 1

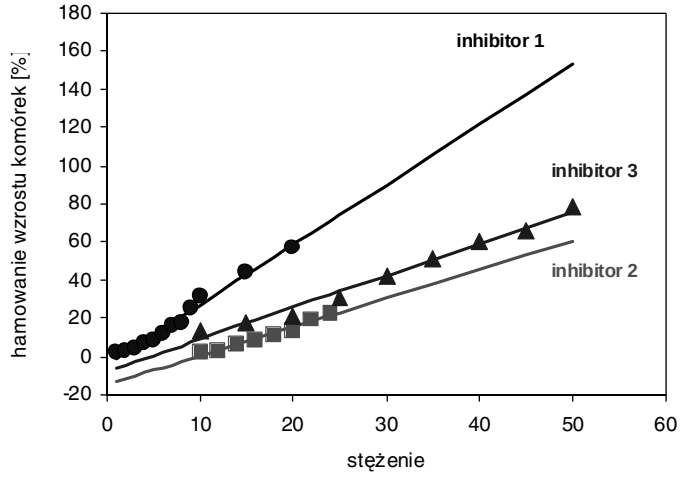
	współczynniki	**błąd stand.**	***t* Studenta**	**istotność**	**dolne 95%**	**górne 95%**
przecięcie (*a*)	–4.400	1.880	–2.34	0.047	–8.74	–0.062
nachylenie (*b*)	3.146	0.303	10.38	0.000	2.45	3.840

Dla inhibitora 2

	współczynniki	błąd stand.	*t* Studenta	istotność	dolne 95%	górne 95%
przecięcie (*a*)	–14.730	1.740	–8.45	0.0002	–18.99	–10.46
nachylenie (*b*)	1.506	0.099	15.22	0.0000	1.26	1.75

Dla inhibitora 3

	współczynniki	błąd stand.	*t* Studenta	istotność	dolne 95%	górne 95%
przecięcie (*a*)	–7.47	2.56	–2.91	0.0225	–13.52	–1.41
nachylenie (*b*)	1.66	0.08	21.16	0.0000	1.47	1.85



Testujemy parę hipotez:
H_0: $b_1 = b_2 = ... = b_k$
H_A: H_0 jest fałszywa

oraz

H_0: $a_1 = a_2 = ... = a_k$
H_A: H_0 jest fałszywa

Wykonujemy obliczenia wstępne:

zmienność	$\sum x^2$	$\sum xy$	$\sum y^2$	*n*	*a*	*b*	SS_{reszt}	df_{reszt}
	A	B	C					
regresja inhibitor 1	385	969	2541	10	–4.4	3.146	102.1402597	8
regresja inhibitor 2	2480	1732	1337	8	–14.73	1.506	127.3935484	6
regresja inhibitor 3	9600	13920	20327	9	–7.47	1.66	143.0000000	7
połączona			27				372.5338000	21
wspólna	12465	16621	24205	27	–0.656	2.104	2042.333253	23
całkowita*	13864	18279	27286	27	–1.779	1.399	3186.04	25

* parametry liczone dla wszystkich danych z trzech grup razem

Zastosujmy najpierw ogólny całościowy test weryfikujący, czy istnieje koincydencja między k równaniami regresji, tzn. sprawdzimy, czy wszystkie współczynniki b_i są sobie równe oraz czy wszystkie współczynniki a_i są sobie równe:

$$F = \frac{\left(\frac{SS_t - SS_p}{2(k-1)}\right)}{\frac{SS_p}{df_p}}$$

(w równaniu tym indeksy dolne „t" dotyczą zmienności całkowitej (*total*), *zaś indeksy „p"* – zmienności połączonej (*pooled*)).

$$SS_t - SS_p = 3186.04 - 372.5338 = 2813.50 \qquad 2(k-1) = 2(3-1) = 4$$

$$SS_p/df_p = 372.5338/21 = 17.73971$$

$$F = \frac{\left(\frac{2813.506}{4}\right)}{17.73971} = 39.65 \; ; \quad \text{z } 2(k-1) \text{ oraz } df_p \text{ stopniami swobody licznika i mianownika.}$$

Ponieważ $F_{0.0001(1),4.21} = 7.83 < F_{dośw} = 39.65$, możemy odrzucić hipotezę zerową i wnioskować, że nie istnieje koincydencja między trzema równaniami regresji, tzn. współczynniki b_1, b_2, b_3 nie są sobie równe oraz/lub współczynniki a_1, a_2, a_3 nie są sobie równe.

Wykonujemy testy szczegółowe i testujemy pierwszą parę hipotez dotyczącą istotności różnic między współczynnikami kierunkowymi (b_i). Wartość statystyki F dla porównania współczynników kierunkowych b wynosi:

$$F = \frac{\left(\frac{SS_c - SS_p}{k-1}\right)}{\frac{SS_p}{df_p}}$$

(indeksy dolne „c" dotyczą zmienności wspólnej (*common*), indeksy *„p"* – zmienności połączonej (*pooled*)).

$$SS_c - SS_p = 1669.8 \quad k-1 = 3-1 = 2 \quad SS_p/df_p = 17.74$$

$$F = \frac{\left(\frac{1669.8}{2}\right)}{17.74} = 47.064 \quad \text{z} \quad k-1 \quad \text{oraz} \quad df_p \text{ stopniami swobody licznika i mianownika.}$$

Ponieważ $F_{0.0001(1),2.21} = 11.2 < F_{dośw} = 47.064$, możemy odrzucić hipotezę zerową i wnioskować, że istnieje istotna różnica między trzema współczynnikami kierunkowymi trzech równań regresji, tzn. że współczynniki b_1, b_2, b_3 nie są sobie równe.

Wartość statystyki F dla porównania współczynników a_i wynosi:

$$F = \frac{\left(\frac{SS_t - SS_c}{k-1}\right)}{\frac{SS_c}{df_c}}$$

$$SS_t - SS_c = 3186.04 - 2042.33 = 1143.71$$

$$k - 1 = 3 - 1 = 2 \qquad SS_c/df_c = 88.8$$

$$F = \frac{\left(\frac{1143.71}{2}\right)}{88.8} = 6.44$$

z $k - 1 = 2$ i $df_c = 23$ stopniami swobody licznika i mianownika.

Ponieważ $F_{0.0001(1),2.23} = 10.8 < F_{dośw} = 6.44$, nie mamy podstaw, aby odrzucić hipotezę zerową i wnioskować, że istnieje istotna różnica między trzema współczynnikami a, czyli że współczynniki a_1, a_2, a_3 nie są sobie równe.

Czytelnik może prześledzić obliczenia do tego przykładu w arkuszu pliku Excela „rozwiazanie zadania 71.xls".

Rozdział 21

Odstające obserwacje

Przykład 72

Zebrano 15 obserwacji dotyczących stężenia HDL w osoczu krwi (mg/100 ml). Jak sprawdzić, czy obserwacje tworzą zwarty zbiór danych i czy nie ma wśród nich wyników odstających?

40 35 43 41 37 15 42 32 34 42 34 58 40 44 38

Porządkujemy wyniki w kolejności od najmniejszego do największego:

15 32 34 34 35 37 38 40 40 41 42 42 43 44 **58**

Ostatni, a szczególnie pierwszy wynik, są najprawdopodobniej odstającymi obserwacjami.

Średnia, mediana i SD dla serii wyników razem z wynikami prawdopodobnie odstającymi wynoszą odpowiednio: $\bar{x} = 38.3$, $Me = 40.0$ i $s = 8.9$. Dla zbioru 13 danych bez wartości skrajnych wynoszą odpowiednio: $\bar{x} = 38.6$, $Me = 40.0$ i $s = 3.9$.

Jeżeli podstawimy zamiast wartości skrajnych najbliżej przylegające do nich liczby:

32 32 34 34 35 37 38 40 40 41 42 42 43 44 **44**

to średnia, mediana i SD będą wynosić odpowiednio: $\bar{x} = 38.5$, $Me = 40.0$ i $s = 4.3$.

Widzimy, że średnia dla „wygładzonych" danych różni się zaledwie o 0.5% od średniej dla wartości początkowych, ale SD jest ponad dwukrotnie niższe. Co by się stało, gdybyśmy zebrali dwukrotnie mniej danych, i jeśli wśród nich znajdowałyby się te same skrajne wartości, jak wpływałyby one na wartość średnią grupy? Nasz nowy hipotetyczny zbiór danych zawiera wartości:

15 32 34 37 40 41 42 44 **58**

zaś zbiór, w którym zastąpiliśmy wartości skrajne najbliżej do nich przylegającymi, wygląda tak:

32 32 34 37 40 41 42 44 **44**

Średnie i odchylenia wynoszą teraz, odpowiednio: 38.1 ± 11.4 oraz 38.4 ± 4.9. Nowa średnia różni się od średniej początkowej o niecały 1%, ale odchylenie jest prawie trzykrotnie niższe, co może mieć znaczenie przy badaniu istotności różnic. Ponieważ wartości skrajne różnią się wyraźnie od pozostałych wyników, zastosujemy teraz testy do zweryfikowania, czy możemy te skrajne wartości traktować jako obserwacje odstające.

Reguła „czterech sigma" zakłada, że obserwacja może być uznana za odstającą w przypadku, gdy różni się od średniej (liczonej bez uwzględnienia wartości „podejrzanej") o istotnie więcej niż 4 wartości odchylenia standardowego. Z charakterystyki rozkładu normalnego wynika, że wartość $z = 4.0$ odpowiada prawdopodobieństwu 99.994%, czyli istnieje około 0.006% szans (tzn. 6 na sto tysięcy), że losowa obserwacja z danej próby będzie miała wartość spoza zakresu $\overline{x} \pm 4s$.

W naszym przypadku:

$$M = \frac{\left|x_i - \overline{X}\right|}{s} = \frac{\left|15 - 38.62\right|}{3.91} = \frac{23.62}{3.91} = 6.04 \quad \text{dla najniższej wartości}$$

oraz $M = \dfrac{\left|x_i - \overline{X}\right|}{s} = \dfrac{\left|58 - 38.62\right|}{3.91} = \dfrac{19.38}{3.91} = 4.957$ dla najwyższej zaobserwowanej wartości.
Ponieważ obie liczby są większe od 4, zarówno najniższy, jak i najwyższy zaobserwowany wynik możemy odrzucić jako obserwacje odstające.

Jeżeli przeprowadzimy obliczenia dla zbioru danych o mniejszej liczebności (*zobacz wyżej*), to uzyskamy:

$$M = \frac{\left|x_i - \overline{X}\right|}{s} = \frac{\left|15 - 38.57\right|}{4.39} = \frac{23.57}{4.39} = 5.37 \quad \text{dla najniższej wartości}$$

oraz $M = \dfrac{\left|x_i - \overline{X}\right|}{s} = \dfrac{\left|58 - 38.57\right|}{4.39} = \dfrac{19.43}{4.39} = 4.426$ dla najwyższej zaobserwowanej wartości.
I tutaj obie liczby są większe od 4, zatem obie obserwacje odrzucamy.

Test Grubbsa polega na uporządkowaniu wszystkich danych w porządku rosnącym oraz obliczeniu średniej i SD na podstawie wszystkich wartości (włączając „podejrzane"). Wartość statystyki T testu Grubbsa* wyliczamy następująco:

$$T = \frac{\overline{X} - x_1}{s} \text{ dla najmniejszej lub } T = \frac{x_n - \overline{X}}{s} \text{ dla największej odstającej obserwacji.}$$

Zależy ona od wielkości próby (n) oraz przyjętego poziomu istotności, który oznacza ryzyko popełnienia błędu przy odrzucaniu odstających obserwacji. Dla najmniejszej (15) i największej wartości (58) w bardziej licznym zbiorze danych.

* Tabele wartości krytycznych może Czytelnik znaleźć w pracy: Grubbs F.E., Beck G.: Extension of sample size and percentage points for significance tests of outlying observations. Technometrics 1972, 14, 847-854.

$$T_1 = \frac{\overline{X} - x_1}{s} = \frac{38.33 - 15}{8.93} = 2.588 \quad \text{oraz} \quad T_n = \frac{x_n - \overline{X}}{s} = \frac{58 - 38.33}{8.93} = 2.203$$

Wartość krytyczna testu Grubbsa dla $n = 15$ przy poziomie istotności $\alpha = 0.05$ wynosi 2.409. Na podstawie wyniku tego testu możemy zatem odrzucić jedynie wartość $x_i = 15$.

Test Q Dixona opiera się na pomiarze proporcji rozstępu między odstającą i następną przylegającą obserwacją do całkowitego rozstępu w próbie. Stosunek taki oblicza się nieco odmiennie dla prób o różnych liczebnościach i zależnie od tego, czy badamy wartości najniższe czy najwyższe. Dla najniższych wartości obserwacji testowanych jako odstające stosujemy równania:

wielkość próby	**równanie**
$3 \le n \le 7$	$\frac{x_2 - x_1}{x_n - x_1}$
$8 \le n \le 10$	$\frac{x_2 - x_1}{x_{n-1} - x_1}$
$11 \le n \le 13$	$\frac{x_3 - x_1}{x_n - x_1}$
$14 \le n \le 25$	$\frac{x_3 - x_1}{x_{n-2} - x_1}$

Z kolei, dla najwyższych wartości obserwacji testowanych, jako odstające stosujemy równania:

wielkość próby	**równanie**
$3 \le n \le 7$	$\frac{x_n - x_{n-1}}{x_n - x_1}$
$8 \le n \le 10$	$\frac{x_n - x_{n-1}}{x_n - x_2}$
$11 \le n \le 13$	$\frac{x_n - x_{n-2}}{x_n - x_2}$
$14 \le n \le 25$	$\frac{x_n - x_{n-2}}{x_n - x_3}$

Obliczoną wartość statystyki Q porównujemy z tablicową wartością krytyczną* dla liczebności n i poziomu istotności α.

Test Q Dixona sprawdza się szczególnie dobrze w przypadku małych zbiorowości danych oraz kiedy pojedyncze obserwacje w próbie są odstające. Test ten posiada mniejszą moc niż test Grubbsa. W przypadku, gdy test Dixona stosujemy do prób, gdzie występuje

* Tabele wartości krytycznych można znaleźć w pracy: Dixon W.J., Massey F.J.: Introduction to Statistical Analysis (table A-8e). McGraw-Hill Book Co., New York 1983.

więcej niż jedna obserwacja odstająca, próba odrzucenia najbardziej skrajnych wartości może prowadzić do nieistotnych wartości statystyki Q, szczególnie w próbach gdzie $n < 10$.

Zastosujmy ten test najpierw do naszego bardziej licznego zbioru (gdzie $n = 15$ obserwacji). Z uwagi na liczebność próby wykorzystamy równania:

$$Q = \frac{x_3 - x_1}{x_{n-2} - x_1} = \frac{34 - 15}{43 - 15} = 0.6786 \qquad \text{oraz} \qquad Q = \frac{x_n - x_{n-2}}{x_n - x_3} = \frac{58 - 43}{58 - 34} = 0.625$$

Ponieważ krytyczna tablicowa wartość Q dla $n = 15$ i $\alpha = 0.05$ wynosi 0.525, oba wyniki możemy odrzucić jako obserwacje odstające. Obliczone przez nas wartości Q są tak wysokie, że najniższą wartość moglibyśmy odrzucić nawet przy poziomie istotności $\alpha = 0.005$, zaś najwyższą przy $\alpha =$ 0.01. Wykonajmy teraz analogiczne obliczenia dla przypadku gdy dysponujemy mniejszą zbiorowością danych ($n = 9$). Wykorzystamy równania:

$$Q = \frac{x_2 - x_1}{x_{n-1} - x_1} = \frac{32 - 15}{44 - 15} = 0.586 \qquad \text{oraz} \qquad Q = \frac{x_n - x_{n-1}}{x_n - x_2} = \frac{58 - 44}{58 - 32} = 0.538$$

Tablicowa wartość Q dla $n = 9$ i $\alpha = 0.05$ wynosi 0.512, toteż oba wyniki możemy odrzucić jako obserwacje odstające.

Przykład 73

Sporządzono krzywą wzorcową fluorescencji znacznika stosowanego do określania stężenia wapnia w komórce.

x_i	y_i	y_c	$y_i - y_c$
2.0	87.1	89.9689	–2.8689
2.5	95.2	93.1572	2.0428
3.0	98.3	96.3456	1.9544
3.5	96.7	99.5339	–2.8339
4.0	100.4	102.7222	–2.3222
4.5	**112.9**	105.9106	**6.9894**
5.0	110.7	109.0989	1.6011
5.5	**108.5**	112.2872	**–3.7872**
6.0	114.7	115.4756	–0.7756
			$\sum = \mathbf{0.0000}$

Oszacowane parametry regresji są nastepujące:

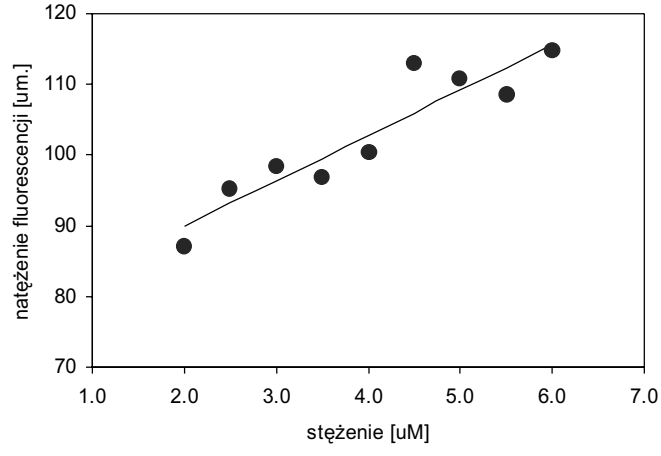
parametry regresji			
korelacja (r)	0.9295	błąd standardowy	3.703
$\mathbf{R^2}$	0.864	liczba obserwacji	9

Analiza wariancji

zmienność	*df*	*SS*	*MS*	*F*	**istotność**
regresji	1	609.93	609.93	44.47	0.0003
resztowa	7	96.01	13.72		
całkowita	8	705.94			

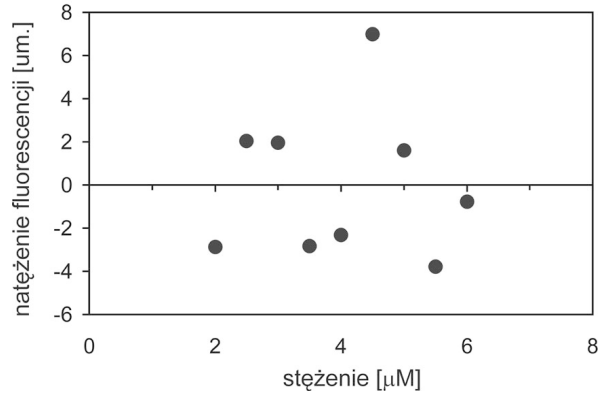
Analiza regresji

	współczynniki	**błąd std.**	***t* Studenta**	***p***	**dolne 95%**	**górne 95%**
przecięcie	77.21556	4.019	19.212	2.6×10^{-7}	67.712	86.719
zmienna x_1	6.376667	0.956	6.669	0.0003	4.116	8.638



Widzimy, że jeden z punktów znacznie odbiega od prostej. Czy ten punkt (4.5; 112.9) można uznać za wynik odstający?

Jaką metodę zastosować? Jeżeli przyjrzymy się wartościom reszt $y_i - y_c$ widzimy, że wynik ten odstaje wyraźnie od innych. Zastosujmy na początku najprostszą ocenę typową dla statystyk jednowymiarowych na podstawie oszacowania przedziału ufności dla reszt. Ponieważ $t_{0.05(2),8} = 2.306$, przedział ufności dla reszt wynosi od –2.663 do +2.663. Zatem stosując ta prostą metodę moglibyśmy uznać, że wynik (4.5; 112.9), jak również wynik (5.5; 108.5) znajdują się poza tym przedziałem i należałoby je uznać za obserwacje odstające. Oczekujemy, że rozkład reszt powinien być równomierny, to znaczy tyle samo punktów pomiarowych powinno leżeć powyżej wartości 0 co poniżej. Na wykresie rozkładu reszt widzimy, że nasz punkt (4.5; 112.9) wyraźnie odstaje od innych obserwacji.



Ponieważ taka wizualna ocena jest mało wiarygodna, możemy przeanalizować rozkład tzw. studentyzowanych reszt obliczonych według równania:

$$t = \frac{y_i - c_c}{\sqrt{MS_{błędu}}}$$

dla $x_i = 4.5$ $y_i = 112.9$ $y_c = bx + a = 6.376667 \text{ x } 4.5 + 77.21556 = 105.91056$

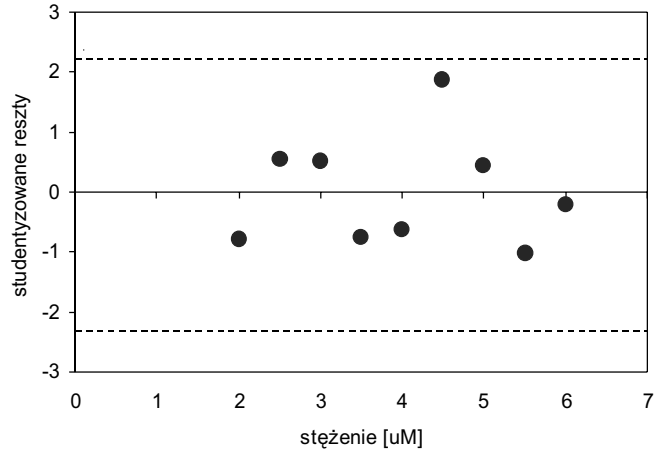
$n - 1 = 8$ $MS_{błędu} = 13.72$

dla $x_i = 5.5$ $y_i = 108.5$ $y_c = bx + a = 6.376667 \text{ x } 5.5 + 77.21556 = 112.28723$

$n - 1 = 8$ $MS_{błędu} = 13.72$

$$t_{4.5} = \frac{y_i - y_c}{\sqrt{MS_{błędu}}} = \frac{112.9 - 105.91056}{\sqrt{13.72}} = \frac{6.98944}{3.7041} = 1.887$$

$$t_{5.5} = \frac{y_i - y_c}{\sqrt{MS_{błędu}}} = \frac{108.5 - 112.28723}{\sqrt{13.72}} = \frac{-3.78723}{3.7041} = -1.022$$



Ponieważ $t_{0.05(2),8} = 2.306$, nie możemy przy poziomie istotności $\alpha = 0.05$ odrzucić wyniku (4.5; 112.9 z $t_{0.05,8} = 1.887$) oraz wyniku (5.5; 108.5 z $t_{0.05,8} = 1.022$) jako obserwacji odstających od innych.

Przykład 74

Czy w zbiorze obserwacji dotyczących frakcji agregatów płytkowych (%):

12.5 12.4 12.9 12.3 12.0

wynik 12.9% można uznać za obserwację istotnie odstającą od reszty?

Porządkujemy wyniki w szeregu rosnącym:

12.0 12.3 12.4 12.5 12.9

Obliczamy średnią i SD:

$$\text{z } 12.9 \quad \bar{x} = 12.42 \quad s = 0.33$$

$$\text{bez } 12.9 \quad \bar{x} = 12.3 \quad s = 0.22$$

Stosujemy test 4σ:

$$M = \frac{\left|x_i - \bar{X}\right|}{s} = \frac{\left|12.3 - 12.9\right|}{0.22} = \frac{0.6}{0.22} = 2.7273$$

Ponieważ 2.7273 <4.0, to nie odrzucamy wartości 12.9.
Dla liczebności $n = 5$ i $\alpha = 0.05$ wartość krytyczna testu Grubbsa wynosi 1.672.

$$T = \frac{x_n - \bar{x}}{s} = \frac{12.9 - 12.42}{0.33} = \frac{0.48}{0.33} = 1.4545$$

jest mniejsze od wartości krytycznej, zatem nie odrzucamy wartości 12.9.
Dla liczebności $n = 5$ i $\alpha = 0.05$ wartość krytyczna testu Q Dixona wynosi 0.642.
Ponieważ:

$$Q = \frac{x_n - x_{n-1}}{x_n - x_1} = \frac{12.9 - 12.5}{12.9 - 12.0} = \frac{0.4}{0.9} = 0.4444$$

jest mniejsze od wartości krytycznej, zatem nie odrzucamy wartości 12.9.

Przykład 75

Zebrano następujące wyniki średniej zawartości DDT mierzone w kolejnych powtórzeniach w jednej próbie:

89.470	85.765	94.673	93.593
89.578	89.954	89.096	90.738
88.975	90.122	89.204	89.711

Czy najwyższa zmierzona wartość (94.673) może być uznana za obserwację odstającą statystycznie istotnie od innych uzyskanych wyników?

Porządkujemy wyniki:

85.765	88.975	89.096	89.204
89.47	89.578	89.711	89.954
90.122	90.738	93.593	94.673

Obliczamy średnią i SD:

$$\text{z } 94.673 \quad \bar{x} = 90.16 \quad s = 2.31$$

$$\text{bez } 94.673 \quad \bar{x} = 89.74 \quad s = 1.90$$

Stosujemy test 4σ:

$$M = \frac{\left|x_i - \overline{X}\right|}{s} = \frac{\left|89.74 - 94.673\right|}{1.90} = \frac{4.933}{1.90} = 2.595$$

Ponieważ 2.595 < 4.0, to nie odrzucamy wartości 94.673.
Dla liczebności $n = 12$ i $\alpha = 0.05$ wartość krytyczna testu Grubbsa wynosi 2.270.

$$T = \frac{x_n - \bar{x}}{s} = \frac{94.673 - 90.16}{2.31} = \frac{4.513}{2.31} = 1.9537$$

jest mniejsze od wartości krytycznej, zatem nie odrzucamy wartości 94.673.
Dla liczebności $n = 12$ i $\alpha = 0.05$ wartość krytyczna testu Q Dixona wynosi 0.546.

Ponieważ:

$$Q = \frac{x_n - x_{n-2}}{x_n - x_2} = \frac{94.673 - 91.738}{94.673 - 88.875} = \frac{2.935}{5.698} = 0.5151$$

jest mniejsze od wartości krytycznej, zatem nie odrzucamy wartości 94.673.

Przykład 76

Badano stężenie znacznika fluorescencyjnego korzystając z dwóch różnych metod: referencyjnej i badanej. Uzyskano wartości przedstawione w tabeli poniżej. Czy wynik dla pary punktów (50; 42) można uznać za odstający od reszty?

metoda referencyjna	metoda badana	metoda referencyjna	metoda badana
30	30.4	80	81.6
40	39.7	90	89.3
50	**42.0**	100	100.1
60	59.1	110	109.7
70	70.8	120	119.4

Policzmy parametry analizy regresji w dwóch wariantach: z uwzględnieniem wyniku potencjalnie odstającego oraz z wyłączeniem tego wyniku.

	z wynikiem odstającym	**bez wyniku odstającego**
n	10.00	9.00
$\sum x$	750.00	700.00
$\sum y$	742.10	700.10
$\sum x^2$	64.50	62000.00
$\sum y^2$	63713.21	61949.21
$\sum xy$	64072.00	61972.00
b	1.02	0.99
a	–2.29	0.79

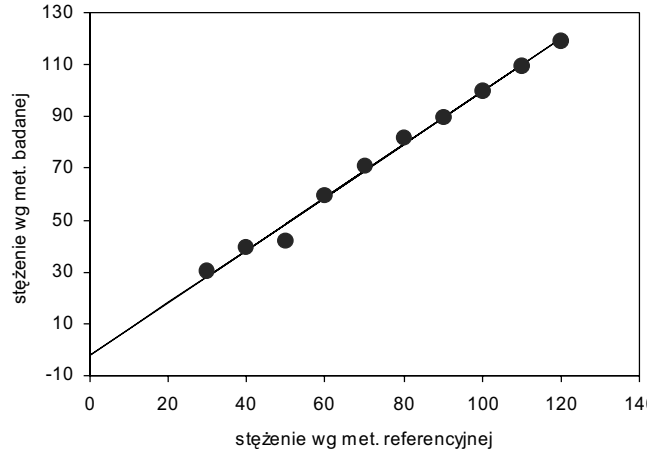
Widzimy, że „podejrzana" obserwacja wpływa jedynie nieznacznie na współczynnik kierunkowy prostej b, ale bardzo istotnie na wartość rzędnej zerowej a. Analiza wariancji i analiza regresji dla danych wyglądają następująco:

Analiza wariancji

zmienność	df	SS	MS	F	**istotność**
regresji	1	8582.28	8582.28	1150.27	6.24×10^{-10}
resztowa	8	59.689	7.461121		
całkowita	9	8641.97			

Analiza regresji

	współczynniki	**błąd std.**	***t* Studenta**	**istotność**	**dolne 95%**	**górne 95%**
a	–2.286	2.415	–0.946	0.372	–7.855	3.284
b	1.02	0.0301	33.916	6.24×10^{-10}	0.951	1.089



Na podstawie współczynników linii regresji wyznaczamy teoretyczne wartości y_c:

x_i	y_i	y_c	$y_i - y_c$
30	30.4	28.313	2.087
40	39.7	38.512	1.188
50	**42.0**	48.712	**–6.712**
60	59.1	58.911	0.189
70	70.8	69.110	1.690
80	81.6	79.310	2.290
90	89.3	89.509	–0.209
100	100.1	99.708	0.392
110	109.7	109.908	–0.208
120	119.4	120.107	–0.707
			$\sum = \mathbf{0.000}$

Możemy uporządkować powyższą tabelę według wartości $y_i - y_c$ od najmniejszej do największej w celu szybkiego policzenia mediany oraz 25-tego i 75-tego percentyla:

x_i	y_i	y_c	$y_i - y_c$
50	**42.0**	48.71152	**–6.712**
120	119.4	120.1073	–0.707
90	89.3	89.50909	–0.209
110	109.7	109.9079	–0.208
60	59.1	58.91091	0.189
100	100.1	99.70848	0.392
40	39.7	38.51212	1.188
70	70.8	69.1103	1.690
30	30.4	28.31273	2.087
80	81.6	79.3097	2.290
			$\sum = \mathbf{0.000}$

Mediana (średnia z 5. i 6. wartości w szeregu) wynosi 0.29, 25-ty percentyl –0.21 (3. wartość), zaś 75-ty percentyl 1.69 (8. wartość w szeregu). Gdybyśmy posłużyli się teraz tymi wartościami do sporządzenia wykresu ramkowego z wąsami (*box and whisker*), to wąsy (wyznaczające umowny zakres wartości) sięgałyby od –2.56 do +3.14. Zatem nasza „wątpliwa" obserwacja wykraczałaby istotnie poza ten zakres i powinniśmy ja odrzucić. Do tych samych danych zastosujmy metodę „studentyzowanego" rozkładu reszt. Dla naszego „podejrzanego" wyniku uzyskamy:

$$t = \frac{y_i - y_c}{\sqrt{MS_{błędu}}} = \frac{42.0 - 48.71152}{\sqrt{7.4611212}} = \frac{-6.71152}{2.731505} = -2.4571$$

Wartość t jest większa od wartości krytycznej wynoszącej $t_{0.05(2),9} = \pm 2.262$, zatem nasz „wątpliwy" wynik (50; 42) możemy odrzucić jako obserwację odstającą.

Rozdział 22

Tablice czteropolowe i wielopolowe

Przykład 77

Spośród 580 osób, 335 zostało zaszczepionych przeciw wirusowi grypy, a pozostałe 245 otrzymało *placebo*. W sumie 146 osób zachorowało na grypę, z czego 28 należało do grupy osób zaszczepionych, a 118 do grupy osób niezaszczepionych. Należy obliczyć, czy szczepienie istotnie zmniejsza częstość zachorowania na grypę.

Liczebności obserwowane

grypa	szczepionka	*placebo*	***razem***
tak	28	118	**146**
nie	307	127	**434**
razem	**335**	**245**	**580**

Naszym pierwszym krokiem będzie policzenie częstości względnych (wartości procentowych) dla każdej komórki tabeli. Częstość osób, które zachorowały w grupie zaszczepionej wynosiła 4.83%, w grupie niezaszczepionej – 20.34%, i w sumie – 25.17%. Już na podstawie wstępnego oszacowania widzimy, że w grupie osób, które dostały *placebo* częstość zachorowań była około 5-krotnie większa.
Nasze hipotezy testowane za pomocą testu chi^2 mają postać:
– H_0: nie ma związku między szczepieniem a zachorowaniem na grypę,
– H_A: jest związek między szczepieniem a zachorowaniem na grypę.

Aby obliczyć wartość statystyki chi^2 musimy policzyć liczebności oczekiwane i porównać je z liczebnościami zaobserwowanymi. W sumie zachorowało 146/580 = 25.17% wszystkich osób badanych i gdyby ryzyko zachorowania było jednakowe wśród osób zaszczepionych, jak wśród osób, które otrzymały *placebo*, to powinniśmy oczekiwać takiej samej proporcji zachorowań w obu grupach, czyli 146/580 x 335 = 84.33 w grupie zaszczepionej i 146/580 x 245 = 61.67 w grupie *placebo*. Podobnie, wśród osób, które uniknęły grypy 434/580 x 335 = 250.67 było zaszczepionych oraz 434/580 x 245 = 183.33 otrzymało *placebo*:

Liczebności oczekiwane

grypa	szczepionka	*placebo*	***razem***
tak	84.33	61.67	**146**
nie	250.67	183.33	**434**
razem	**335**	**245**	**580**

Obliczamy wartość statystyki chi^2 według równania:

$$\chi^2 = \sum \frac{(f_{obs} - f_{oczek})^2}{f_{oczek}}$$

$$\chi^2 = \frac{(28 - 84.33)^2}{84.33} + \frac{(118 - 61.67)^2}{61.67} + \frac{(307 - 250.67)^2}{250.67} + \frac{(127 - 183.33)^2}{183.33} =$$

$$= 37.63 + 51.45 + 12.66 + 17.31 = 119.05$$

Wartość $\chi^2 = 119.05$ jest o wiele większa od wartości tablicowej $\chi^2_{0.001,1} = 10.83$, czyli prawdopodobieństwo, że wyższe częstości zachorowań na grypę wśród osób nie zaszczepionych są dziełem przypadku wynosi mniej niż 0.1%. Z prawdopodobieństwem ponad 99.9% możemy odrzucić hipotezę zerową i wnioskować, że szczepionka jest skuteczna w zmniejszaniu ryzyka zachorowania na grypę.

Przykład 78

Dla tych samych danych należy policzyć wartość statystyki chi^2 z uwzględnieniem poprawki na ciągłość Yatesa.

Uwzględniając poprawkę mamy:

$$\chi^2 = \sum \frac{\left(|f_{obs} - f_{oczek}| - \frac{1}{2}\right)^2}{f_{oczek}}$$

$$\chi^2 = \frac{\left(28 - 84.33 - \frac{1}{2}\right)^2}{84.33} + \frac{\left(118 - 61.67 - \frac{1}{2}\right)^2}{61.67} + \frac{\left(307 - 250.67 - \frac{1}{2}\right)^2}{250.67} + \frac{\left(127 - 183.33 - \frac{1}{2}\right)^2}{183.33} =$$

$$= 36.96 + 50.54 + 12.43 + 17.00 = 116.93$$

Wartość statystyki chi^2 z uwzględnieniem poprawki jest nieco niższa, ale istotność wnioskowania jest nadal bardzo wysoka (>99.9%, $p < 0.001$).

Przykład 79

Dla danych z przykładu 74 policz wartość statystyki z uwzględniając poprawkę na ciągłość w celu zweryfikowania hipotez:

- H_0: proporcje osób, które zachorowały na grypę wśród szczepionych i nie szczepionych są równe
- H_A: proporcje osób, które zachorowały na grypę wśród szczepionych i nie szczepionych *nie* są równe

Wśród 335 osób, które zostały zaszczepione 28 zachorowało na grypę, co stanowi 28/335 = 0.0836 czyli 8.36%. Z kolei wśród 245 osób, które nie zostały zaszczepione (dostały *placebo*) 118 osób zachorowało na grypę, co stanowi 0.4816 czyli 48.16%.

Całkowita proporcja osób, które zachorowały wynosi:

$$p = \frac{28+118}{335+245} = \frac{146}{580} = 0.2517 \quad \text{czyli} \quad 25.17\%, \quad p_1 = 0.0836, \quad p_2 = 0.4816$$

Wartość statystyki z z poprawką na ciągłość obliczymy z równania:

$$z = \frac{|p_1 - p_2| - [1/(2n_1) + 1/(2n_2)]}{\sqrt{[p(1-p)(1/n_1 + 1/n_2)]}}$$

$$z = \frac{|0.0836 - 0.4816| - [1/(670) + 1/(490)]}{\sqrt{[0.2517(1-0.2517)(1/335 + 1/245)]}} = \frac{0.398 - 0.003533}{\sqrt{(0.18835)(0.00707)}} = \frac{0.39447}{0.0365} = 10.8099$$

Skoro $z = 10.81$ to $z^2 = 116.85$. Zatem wartość z^2 jest bardzo bliska obliczonej w poprzednim przykładzie wartości $\chi^2 = 116.93$.

Korzystając z „udogodnień" testu normalnego możemy także oszacować granice przedziału ufności dla różnicy między proporcjami:

$$CI\,95\% = (p_1 - p_2) \pm 1.96\sqrt{[p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2]}$$

$$CI\,95\% = (-0.398) \pm 1.96\sqrt{[0.0836(1-0.0836)/335 + 0.4816(1-0.4816)/245]} =$$

$$= (-0.398) \pm (1.96 \text{ x } 0.03532) = -0.329 \text{ do } -0.467$$

Czyli, prawdopodobieństwo zachorowania na grypę jest mniejsze o 32.9% do 46.7% wśród osób zaszczepionych w porównaniu z osobami, które otrzymały *placebo*.

Przykład 80

Badano preferencje trzech gatunków chruścików do zasiedlania zbiorników z wodą stojącą, oraz zbiorników z wolnym i szybkim przepływem wody. Należy ocenić, czy reżim wody ma wpływ na zasiedlanie zbiornika przez każdy z gatunków.

typ zbiornika	gatunek 1	gatunek 2	gatunek 3	***razem***
woda stojąca	20	32	18	**70**
wolny przepływ	18	20	12	**50**
szybki przepływ	12	8	10	**30**
razem	**50**	**60**	**40**	**150**

W celu wstępnej oceny różnic między komórkami obliczamy częstości względne. Przedstawiają się one następująco:

typ zbiornika	gatunek 1	gatunek 2	gatunek 3	***razem***
woda stojąca	40	53.33	45	**46.67**
wolny przepływ	36	33.33	30	**33.33**
szybki przepływ	24	13.33	25	**20.00**
razem	**100**	**100.00**	**100**	**100.00**

Widzimy, że każdy z gatunków najchętniej zasiedla wody stojące, a najmniej chętnie wody o szybkim nurcie.

Obliczamy liczebności oczekiwane. Możemy to zrobić według schematu:

$$\frac{70}{150}\times 50 = 23.33\,, \qquad \frac{50}{150}\times 50 = 16.67\,, \qquad \frac{30}{150}\times 50 = 10.0$$

tak jak dla gatunku 1, lub wykorzystując równanie:

$$f_{oczek(11)} = \frac{(suma\ kolumny)_1 \times (suma\ wiersza)_1}{suma\ całkowita} = \frac{50\times 70}{150} = 23.33$$

Policzone w ten sposób liczebności oczekiwane dla każdej komórki tabeli wynoszą:

typ zbiornika	gatunek 1	gatunek 2	gatunek 3	***razem***
woda stojąca	23.33	28.00	18.67	**70**
wolny przepływ	16.67	20.00	13.33	**50**
szybki przepływ	10.00	12.00	8.00	**30**
razem	**50.00**	**60.00**	**40.00**	**150**

Obliczamy wartość statystyki chi^2:

$$\chi^2 = \sum \frac{(f_{obs} - f_{oczek})^2}{f_{oczek}} =$$

$$= (20 - 23.33)^2/23.33 + (32 - 28)^2/32 + (18 - 18.67)^2/18.67 + (18 - 16.67)^2/16.67 +$$

$$+ (20 - 20.0)^2/20.0 + (12 - 13.33)^2/13.33 + (12 - 10.0)^2/10.0 + (8 - 12.0)^2/12.0 +$$
$$+ (10 - 8.0)^2/8.0 = 3.531$$

Liczba stopni swobody wynosi:

$$d.f. = (r-1)\times(c-1) = (3-1) \text{ x } (3-1) = 2 \text{ x } 2 = 4$$

Wartość krytyczna χ^2 odczytana w tablicach wynosi 5.39 dla $\alpha = 0.05$, zatem nie mamy podstaw do odrzucenia hipotezy zerowej: nie ma istotnych różnic w preferencjach różnych gatunków chruścików w stosunku do źródeł wody o różnym reżimie.

Przykład 81

Oceniano ryzyko zakażenia wirusowym zapaleniem wątroby typu B wśród różnych grup zawodowych:

WZW B	laboranci	biochemicy	lekarze	nauczyciele	***razem***
dodatni	52	38	15	16	**121**
ujemny	21	17	28	44	**110**
razem	**73**	**55**	**43**	**60**	**231**

Czy ryzyko infekcji jest różne u przedstawicieli różnych zawodów?

Do analizy liczebności wykorzystamy wzór do porównania 4 proporcji, z których każda jest przypisana do jednej grupy zawodowej:

$$\chi^2 = \frac{N^2\left[\sum\left(r^2/n\right) - R^2/N\right]}{R(N-R)}$$

Liczymy składniki licznika:

$$\sum\left(r^2/n\right) = 52^2/73 + 38^2/55 + 15^2/43 + 16^2/60 = 37.04 + 26.25 + 5.23 + 4.27 = 72.79$$

$$R^2/N = 121^2/231 = 63.38$$

$$\chi^2 = \frac{231^2(72.79 - 63.68)}{121\times110} = \frac{53361\times9.414}{13310} = 37.74, \quad d.f. = 4-1 = 3$$

Ponieważ $\chi^2_{dośw} > \chi^2_{0.001,3} = 16.27$, możemy wnioskować z prawdopodobieństwem przynajmniej 99.9%, że ryzyko zakażenia jest różne dla różnych grup zawodowych.

Test dokładny Fishera

Przykład 82

W poniższej tabeli przedstawiono wyniki porównania skuteczności dwóch różnych metod zapobiegania krwawieniom u osób z hemofilią:

krwawienia	metoda A	metoda B	*razem*
tak	1	3	**4**
nie	12	9	**21**
razem	**13**	**12**	**25**

Czy możemy uznać, że istnieje zależność między zastosowaną metodą a częstością krwawień?

Widzimy, że u 1 z 13 pacjentów wystąpiły krwawienia po zastosowaniu metody A (7.7%) w porównaniu do 3/12 (25%) u pacjentów leczonych metodą B.
Ponieważ liczebności oczekiwane w wierszu 1 są mniejsze od 5:

krwawienia	metoda A	metoda B	*razem*
tak	2.08	1.92	**4**
nie	10.92	10.08	**21**
razem	**13.00**	**12.00**	**25**

do obliczeń nie możemy wykorzystać testu chi^2.

Posłużymy się testem dokładnym Fishera i policzymy dokładne prawdopodobieństwo tablicy 2 x 2:

$$P_{(2\times2)} = \frac{e!\,f!\,g!\,h!}{n!\,a!\,b!\,c!\,d!} =$$

$$\frac{(4!)(21!)(13!)(12!)}{(25!)(1!)(3!)(12!)(9!)} = \frac{4\times(3!)(21!)\times13\times(12!)\times12\times11\times10\times(9!)}{25\times24\times23\times22\times(21!)(1!)(3!)(12!)(9!)} =$$

$$= \frac{4\times13\times12\times11\times10}{25\times24\times23\times22} = 0.2261$$

Jeżeli mamy przetestować hipotezę zerową mówiącą, że nie ma różnic między dwoma metodami leczenia, czyli że nie istnieje zależność między metodą leczenia a częstością występowania krwawień, możemy sobie postawić pytanie: jakie jest prawdopodobieństwo, że liczebności w poszczególnych komórkach ułożą się tak nierównomiernie (lub gorzej) jak to obserwujemy w analizowanej tabeli czteropolowej? Dla małych liczebności całkowitych n prawdopodobieństwo takie można policzyć dokładnie przez zliczenie wszystkich możliwych tabel, które można byłoby skonstruować na podstawie liczebności brzegowych. W ten sposób dokładny test Fishera oblicza dokładne prawdopodobieństwo przy hipotezie zerowej mówiącej o uzyskaniu rozkładu liczebności w komórkach takiego samego jak obserwowany lub jeszcze bardziej nierównomiernego. W kombinacjach tych uwzględnia się zarówno prawdopodobieństwa jedno- jak i obustronne.

Aby przetestować prawdziwość hipotezy zerowej mówiącej, że nie ma różnicy między metodami leczenia, powinniśmy obliczyć nie tylko prawdopodobieństwo wystąpienia liczebności obserwowanych w tabeli, ale także prawdopodobieństwo zdarzeń bardziej skrajnych, które mogłyby pojawić się przez czysty przypadek. Dla naszego przykładu, istnieje 5 możliwych układów tabel liczebności z takimi samymi liczebnościami brzegowymi jak w tabeli analizowanej.

Tabela a)

krwawienia	metoda A	metoda B	*razem*
tak	0	4	**4**
nie	13	8	**21**
razem	**13**	**12**	**25**

Tabela b)

krwawienia	metoda A	metoda B	*razem*
tak	1	3	**4**
nie	12	9	**21**
razem	**13**	**12**	**25**

Tabela c)

krwawienia	metoda A	metoda B	*razem*
tak	2	2	**4**
nie	11	10	**21**
razem	**13**	**12**	**25**

Tabela d)

krwawienia	metoda A	metoda B	*razem*
tak	3	1	4
nie	10	11	**21**
razem	**13**	**12**	**25**

Tabela e)

krwawienia	metoda A	metoda B	*razem*
tak	4	0	**4**
nie	9	12	**21**
razem	**13**	**12**	**25**

Tabela b) reprezentuje nasz przypadek. Dla poszczególnych tabel wartości dokładnego prawdopodobieństwa wynoszą:

dla tabeli a): $p = 0.03913$
dla tabeli b): $p = 0.2261$
dla tabeli c): $p = 0.407$
dla tabeli d): $p = 0.2713$
dla tabeli e): $p = 0.0565$

Bardziej skrajne warianty liczebności komórek tabeli to te, którym odpowiadają mniejsze prawdopodobieństwa, czyli tabela a) i tabela e). Zatem całkowite prawdopodobieństwo wynosi: 0.2261 + 0.0391 + 0.0565 = 0.3217, czyli różnica między metodami leczenia jest nieistotna.

Zauważmy, że zastosowaliśmy tutaj metodę rachunkową oszacowania istotności różnic jako sumy prawdopodobieństwa wystąpienia liczebności takich jak te, obserwowane w tabeli plus prawdopodobieństw wystąpienia liczebności bardziej skrajnych (czyli mniej prawdopodobnych tabel).

Spróbujmy dla tych samych wyników policzyć teraz istotność jako podwojoną sumę prawdopodobieństwa wystąpienia liczebności takich jak te obserwowane w tabeli plus prawdopodobieństw wystąpienia liczebności bardziej skrajnych (czyli mniej prawdopodobnych tabel) przy zachowaniu tego samego kierunku zmian, co obserwowany w tabeli. Mnożnik 2 zastosowany jest po to, aby skorygować szacowaną wartość prawdopodobieństwa o prawdopodobieństwa wariantów liczebności obserwowanych także w przeciwnym kierunku. Jest to zabieg analogiczny do tego, który wykonujemy w przypadku testu normalnego, gdy obliczamy istotność testu obustronnego poprzez podwojenie istotności testu jednostronnego.

W naszym przypadku interesują nas prawdopodobieństwa obliczone na podstawie rozkładu liczebności w tabelach:
tabela a) to nasza mniej prawdopodobna (skrajna) tabela o rozkładzie liczebności w tym samym kierunku co w tabeli z liczebnościami obserwowanymi:

krwawienia	metoda A	metoda B	*razem*
tak	0	4	**4**
nie	13	8	**21**
razem	**13**	**12**	**25**

tabela b) to nasza właściwa tabela:

krwawienia	metoda A	metoda B	*razem*
tak	1	3	**4**
nie	12	9	**21**
razem	13	12	25

Tak szacowana istotność wynosi 2 x (0.0391 + 0.2261) = 0.5304.

Chociaż żadna z tych dwóch metod obliczeń nie ma przewagi nad drugą, to drugie rachunkowe podejście jest prostsze i szybsze w wykonaniu. Zauważmy jednak, że obie metody dają różne wyniki prawdopodobieństwa różnic. W praktyce nie ma to jednak znaczenia, gdyż niezależnie od wartości bezwzględnej wyniku nieprzypadkowa istotność różnic między metodami jest i tak mało prawdopodobna.

Testy do pomiarów sparowanych

Przykład 83

W poniższej tabeli przedstawiono wyniki porównania dwóch metod badania skuteczności kwasu acetylosalicylowego (ASA) w hamowaniu funkcji płytek krwi: a) oznaczania agregacji we krwi pełnej (WBA) oraz *b*) oznaczania czasu okluzji (CT) w analizatorze funkcji płytek PFA-100™:

wynik	WBA	CT	*razem*
+	253	215	**468**
–	86	124	**210**
razem	**339**	**339**	**678**

Każdą metodą wykonano oznaczenia dla 339 próbek (pochodzących od 339 pacjentów) i zarejestrowano 75% wyników dodatnich dla metody WBA oraz 63% wyników dodatnich dla metody CT. Na podstawie wartości statystyki chi^2 (9.96 dla df=1) moglibyśmy orzec, że dwie metody różnią się istotnie pod względem wykrywania skuteczności ASA w hamowaniu funkcji płytek krwi. Zauważmy jednak, że takie zestawienie tabeli jest niepoprawne, gdyż nie uwzględnia sparowania danych, to znaczy tego, że te same 339 prób było oznaczone jedną i drugą metodą. Prób było 339, a nie 339 (w metodzie WBA) plus 339 (w metodzie CT) czyli 678. Prawidłowe zestawienie tabeli dla testu sparowanego powinno wyglądać tak:

		CT		
		+	–	*razem*
WBA	+	196	57	**253**
	–	19	67	**86**
	razem	**215**	**124**	**339**

196 wyników było dodatnich w obu metodach oraz 67 wyników było ujemnych w obu metodach. Wskazuje to, że znacznie częściej obserwowano wynik dodatni niż wynik ujemny po podaniu ASA, niezależnie od zastosowanej metody badania funkcji pytek. Te 263 wyniki (196 + 67) nie dają nam jednak żadnej informacji na temat tego, która z metod jest lepsza w wykrywaniu wrażliwości płytek na działanie ASA. To te pozostałe 76 wyników (339 – 263) decyduje o równocenności lub nierównocenności obu metod diagnostycznych. Spośród tych 76 wyników 57 (na 339 = 16.8%) było dodatnich jedynie w badaniu metodą WBA (ale nie metodą CT), zaś 19 (na 339 = 5.6%) było dodatnich jedynie w badaniu metodą CT (ale nie metodą WBA).

Oczywiście nie powinniśmy oczekiwać idealnej 100% zgodności między metodami, nawet w przypadku, gdyby nie było różnic między dwoma porównywanymi metodami, gdyż analizowaliśmy 339 różnych próbek pochodzących od 339 różnych pacjentów. Powinniśmy jednak spodziewać się podobnej częstości przypadków niezgodności (niespójności wyników) dla obu porównywanych metod, to znaczy około 50% dodatnich wyników jedynie dla metody WBA i około 50% wyników dodatnich jedynie dla metody CT. W analizowanym przypadku, wartości te wynosiły odpowiednio, 16.8% i 5.6%. Aby ocenić, czy taka różnica jest istotna statystycznie, powinniśmy porównać, czy proporcja 57/76 lub proporcja 19/76 są istotnie różne od 0.5. Możemy to zrobić stosując test normalny dla proporcji z poprawką na ciągłość:

$$z = \frac{|p - \pi| - 1/(2n)}{\sqrt{[\pi(1-\pi)/n]}}$$

$$z = \frac{|57/76 - 0.5| - 1/(2\times 76)}{\sqrt{[0.5(1-0.5)/76]}} = 4.24 \qquad \text{lub} \qquad z = \frac{|19/76 - 0.5| - 1/(2\times 76)}{\sqrt{[0.5(1-0.5)/76]}} = 4.24$$

Ponieważ $z_{0.0001} = 3.99$, możemy wnioskować z prawdopodobieństwem ponad 99.9%, że obie metody nie są równocenne i że metoda WBA jest istotnie lepsza.

Przykład 84

Dla tych samych wyników zastosujmy teraz test McNemara w celu określenia czy obie porównywane metody diagnostyczne dają podobne wyniki monitorowania skuteczności hamowania czynności płytek krwi przez ASA.

Wykorzystamy równanie opisujące wartość statystyki chi^2 uwzględniającej liczbę par niezgodnych dla pomiarów sparowanych:

$$\chi^2_{par} = \frac{(|b-c|-1)^2}{b+c}$$

W naszym przykładzie

$$b = 57 \quad \text{i} \quad c = 19.$$

Zatem

$$\chi^2_{par} = \frac{(|57-19|-1)^2}{57+19} = \frac{37^2}{76} = 18.01, \quad \text{dla } df = 1 \qquad p < 0.0001$$

Możemy zauważyć, że $\chi^2_{\text{par}} = z^2 = 18.01$: oba testy są matematycznie równoważne.

Przykład 85

Dla danych z przykładu 83 należy policzyć różnicę między dwoma proporcjami, błąd standardowy oraz przedział ufności.

Proporcje wyników dodatnich dla każdej z metod wynoszą:

WBA: 253/339 = 0.7463
CT: 215/339 = 0.6342

Różnica między proporcjami wynosi:

$$0.7463 - 0.6342 = 0.1121$$

Wykorzystując równanie:

$$\textit{różnica} = \frac{b-c}{n}$$

obliczymy, że różnica między proporcjami wynosi:

$$\frac{57-19}{339} = 0.1121$$

Widzimy zatem, że te dwa sposoby szacowania różnicy proporcji są równoważne.

Błąd standardowy wynosi:

$$SE = \frac{\sqrt{b+c}}{n} = \frac{\sqrt{57+19}}{339} = \frac{8.718}{339} = 0.02572$$

$$95\%CI = \frac{(b-c)}{n} \pm z' \frac{\sqrt{(b+c)}}{n} = \frac{(57-19)}{339} \pm 1.96 \frac{\sqrt{(57+19)}}{339} = 0.1121 \pm 1.96 \text{ x } 0.02572$$

$$95\%CI = 0.0617, \quad 0.1625$$

Przedział ufności (95%) dla różnicy między tymi proporcjami wynosi od 0.0617 do 0.1625.

Tabele zbiorcze 2 x 2

Przykład 86

W tabeli poniżej przedstawiono wyniki częstości występowania zespołu zaburzeń oddychania (RDS) u noworodków donoszonych i przedwcześnie urodzonych (dane zostały zestawione osobno dla dziewczynek i chłopców):

a) dziewczynki

RDS	donoszone	wcześniaki	***razem***
tak	11 (11.1%)	22 (14.4%)	**33**
nie	88	131	**219**
razem	**99**	**153**	**252**
$\chi^2 = 0.564$	$df = 1$	ns	

b) chłopcy

RDS	donoszone	wcześniaki	***razem***
tak	48 (47.5%)	36 (73.5%)	**84**
nie	53	13	**66**
razem	**101**	**49**	**150**
$\chi^2 = 9.01$	$df = 1$	$p < 0.001$	

c) razem

RDS	donoszone	wcześniaki	***razem***
tak	59 (29.5%)	58 (28.7%)	**117**
nie	141	144	**285**
razem	**200**	**202**	**402**
$\chi^2 = 0.03$	$df = 1$	ns	

Zarówno w przypadku chłopców, jak i dziewczynek, u wcześniaków RDS występuje częściej niż u dzieci donoszonych, ale różnica taka znika kiedy połączymy obie grupy. Jest to wynikiem działania dwóch czynników:

a) po pierwsze, częstość występowania RDS nie jest jednakowa u chłopców i dziewczynek: jest znacznie wyższa u chłopców,
b) po drugie, proporcja chłopców i dziewczynek wśród noworodków donoszonych i wcześniaków jest inna: wśród wcześniaków było 49/202 = 24% chłopców, natomiast wśród noworodków donoszonych 101/200 = 51% chłopców.

Płeć jest tutaj zmienną uwikłaną (towarzyszącą) związaną zarówno z częstością RDS, jak i ze zjawiskiem donoszenia ciąży. Pomijanie wpływu płci na rozkład liczebności fałszuje wyniki analizy, jak to pokazuje tabela liczebności połączonych. Możemy zatem wnioskować, że analiza liczebności zsumowanych dla obu płci razem zaciemnia i maskuje wpływ zjawiska wcześniactwa na występowanie RDS u noworodków.

Przykład 87

Dla danych z poprzedniego przykładu należy obliczyć wartość sumarycznej statystyki chi^2 dla układu zawierającego dwie podgrupy płci.

Posługujemy się równaniem obliczania statystyki chi^2 testu Mantela-Haenszela z uwzględnieniem poprawki na ciągłość:

$$\chi^2_{MH} = \frac{\left(\left|\sum a - \sum f_{oczek(a)}\right| - 0.5\right)^2}{\sum V_a} \qquad d.f. = 1$$

gdzie

a jest liczebnością obserwowaną,
$f_{oczek(a)} = eg/n$ jest liczebność oczekiwaną dla a,
$V_a = efgh/[n^2(n-1)]$ oznacza zmienność liczebności a.

Dla danych występowania RDS u noworodków mamy:

podgrupa	**a**	**E_a**	**V_a**	
chłopcy	36	27.44	84 x 66 x 101 x 49/(150^2 x 149)	**8.1841**
dziewczynki	22	20.04	33 x 219 x 99 x 153/(250^2 x 249)	**6.8677**
razem	**58**	**47.48**		**15.0518**

$$\chi^2_{MH} = \frac{\left(|58 - 47.48| - 0.5\right)^2}{15.0518} = \frac{(10.02)^2}{15.0518} = 6.68, \quad df = 1 \quad p < 0.01$$

Aby upewnić się czy nie są naruszone ograniczenia zastosowania testu Mantela-Haenszela, stosujemy tzw. „regułę 5". Mówi ona, że suma wartości minimalnych (e, g), oraz suma wartości maksymalnych (0, $g - f$) powinny się różnić od sumy liczebności oczekiwanych przynajmniej o 5:

podgrupa	min (*e*, *g*)	*g* – *f*	max (0, *g* – *f*)
chłopcy	84	35	35
dziewczynki	33	–120	0
razem	**117**		**35**

Ponieważ obie sumy: 117 i 35 różnią się od 47.48 o więcej niż 5, założenia stosowania testu chi^2 Mantela-Haenszela nie są naruszone.

Badanie trendu

Przykład 88

Poniżej zestawiono rozkład liczebności pacjentów z łagodną i ostrą postacią choroby wieńcowej, u których badano wpływ wzrastających dawek aggrastatu (zachowując odpowiednie przerwy między stosowaniem różnych dawek leku) na osłabienie reaktywności płytek krwi, monitorowanej jako wydłużanie czasu okluzji w analizatorze PFA-100™. Czy na podstawie poniższych wyników można stwierdzić, że wzrost dawki aggrastatu powoduje stopniowe obniżanie reaktywności płytek krwi u pacjentów z łagodną postacią choroby wieńcowej?

	dawka leku			
postać ch-w	**mała**	**średnia**	**duża**	*razem*
łagodna	145 (18%)	297 (37%)	354 (44%)	**796**
ostra	156	189	168	**513**
razem	**301**	**486**	**522**	**1309**
punktacja grupy	1	2	3	

Widzimy, że ze wzrostem dawki leku liczba pacjentów z łagodną postacią choroby wieńcowej, u których zanotowano wydłużenie czasu okluzji (wskazujące na osłabienie reaktywności płytek), wzrasta. (Trend taki nie jest tak wyraźny w grupie z ostrą postacią choroby wieńcowej). Dla grupy z łagodną chorobą wieńcową przyjmujemy zatem kierunek trendu od małej do dużej dawki leku i w taki sposób numerujemy kolumny tabeli.

Obliczamy wartość statystyki chi^2 dla trendu według równania:

$$\chi^2_{trend} = \frac{(|A| - 0.5)^2}{B}$$

$$A = \sum (rx) - \frac{R}{N} \sum (nx) \quad \text{i} \quad B = \frac{R(N-R)}{N^2(N-1)} \left[N \sum (nx^2) - (\sum nx)^2 \right]$$

gdzie:

$\sum (rx) = 145 \text{ x } 1 + 297 \text{ x } 2 + 354 \text{ x } 3 = 1801$

$\sum (nx) = 301 \text{ x } 1 + 486 \text{ x } 2 + 522 \text{ x } 3 = 2839$

$\sum (nx^2) = 301 \text{ x } 1 + 486 \text{ x } 4 + 522 \text{ x } 9 = 6943$

$$R = 796, \quad N = 1309, \quad N - R = 513, \quad R/N = 0.608$$

$$A = \sum(rx) - \frac{R}{N}\sum(nx) = 1801 - \frac{796}{1309}(2839) = 74.61$$

$$B = \frac{R(N-R)}{N^2(N-1)}\left[N\sum(nx^2) - (\sum nx)^2\right] = \frac{796 \times 513}{1309^2(1308)}\left[1309 \times 6943 - 2839^2\right] = 187.384$$

$$\chi^2_{trend} = \frac{(|A| - 0.5)^2}{B} = \frac{(|74.61| - 0.5)^2}{187.384} = 29.31$$

Z prawdopodobieństwem ponad 99.9% możemy wnioskować, że obserwowany trend jest istotny, czyli że wzrost dawki aggrastatu powoduje stopniowe obniżanie reaktywności płytek krwi u pacjentów z łagodną postacią choroby wieńcowej.

Przykład 89

Czy podobny trend występuje w grupie pacjentów z ostrą postacią choroby wieńcowej?

	dawka leku			
postać ch-w	**mała**	**średnia**	**duża**	*razem*
łagodna	145	297	354	**796**
ostra	156 (30%)	189 (37%)	168 (33%)	**513**
razem	**301**	**486**	**522**	**1309**
punktacja grupy	1	3	2	

Dla grupy pacjentów z ostrą postacią choroby wieńcowej wartości statystyk będą wynosić:

$\sum(rx) =$ 156 x 1 + 189 x 3 + 168 x 2 = 1059

$\sum(nx) =$ 301 x 1 + 486 x 3 + 522 x 2 = 2803

$\sum(nx^2) =$ 301 x 1 + 486 x 9 + 522 x 4 = 6763

$$R = 513, \quad N = 1309, \quad N - R = 796, \quad R/N = 0.392$$

$$A = \sum(rx) - \frac{R}{N}\sum(nx) = 1059 - \frac{513}{1309}(2803) = -39.502$$

$$B = \frac{R(N-R)}{N^2(N-1)}\left[N\sum(nx^2) - (\sum nx)^2\right] = \frac{513 \times 796}{1309^2(1308)}\left[1309 \times 6763 - 2803^2\right] = 181.4615$$

$$\chi^2_{trend} = \frac{(|A|)^2}{B} = \frac{(|-39.502|)^2}{181.4615} = 8.599$$

W równaniu tym nie uwzględniamy poprawki na ciągłość, gdyż numery wyznaczające kolejność trendu nie odpowiadają kolejności kolumn tabeli.

Na podstawie wartości $\chi^2_{0.005,1} < \chi^2_{trend} = 8.6$ możemy wnioskować, że również w tej grupie występuje trend obniżania reaktywności płytek krwi wraz ze wzrostem dawki aggrastatu. Zauważmy jednak, że chociaż trend w tej grupie jest jedynie częściowo zachowany (najmniejsza liczebność dla małej dawki leku, ale największa dla średniej, a nie dla wysokiej dawki leku) istotność statystyki testu chi^2 jest zachowana; jest ona jednak prawie 4-krotnie niższa niż wartość liczona dla grupy z łagodną postacią choroby wieńcowej.

Rozdział 23

Zastosowania wybranych analiz wielowymiarowych

Analiza funkcji dyskryminacyjnej

Przykład 90

Zbiór ***anal-dyskryminacyjna.xls*** zawiera dane wybranych parametrów biochemicznych oraz klinicznych zebranych dla pacjentów z chorobą niedokrwienną serca. Wśród pacjentów tych znajdowali się tacy chorzy, u których nie wystąpił zawał mięśnia sercowego, pacjenci z zawałem niepełnościennym oraz pacjenci z pełnościennym zawałem mięśnia sercowego. Czy na podstawie analizy wieloparametrowej możemy stwierdzić, że badane parametry kliniczne i biochemiczne istotnie rozróżniają te trzy grupy pacjentów? Które parametry charakteryzują się największą zdolnością dyskryminacji między badanymi grupami pacjentów?

Analizę funkcji dyskryminacyjnej przeprowadzimy przy wykorzystaniu pakietu statystycznego Statistica ver 5.3 (*StatSoft*).

W celu określenia które parametry charakteryzują się największą zdolnością dyskryminacji między badanymi grupami pacjentów zastosowano standardowy model analizy funkcji dyskryminacyjnej (z usuwaniem przypadków z brakującymi polami), obejmujący następujące zmienne:

HDL	cholesterol frakcji HDL osocza [mg/100 ml]
LDL	cholesterol frakcji LDL osocza [mg/100 ml]
TG	triglicerydy osocza [mg/100 ml]
PAI-1	PAI-1 osocza [ng/ml]
Fg	fibrynogen osocza [g/L]
Plt	miano płytek [G/L]
ciśnienie skurczowe	mmHg
ciśnienie rozkurczowe	mmHg
BMI	wskaźnik masy ciała
wiek	

Zmienną grupującą (kategoryzującą) jest rodzaj zawału (bez zawału, zawał niepełnościenny, zawał pełnościenny).

Za kryterium włączania/wykluczania zmiennych przyjęto wartości statystyki *F*, wartości współczynnika tolerancji (1-R^2) oraz wartości statystyki lambda Wilks'a. Współczynnik *F* wskazuje, które zmienne przyczyniają się w sposób najbardziej istotny do wyraźnej dyskryminacji między grupami. Niskie wartości współczynnika tolerancji są wskazaniem zbędności danego parametru w analizie (ponieważ sugerują jego wysoką korelację z innymi parametrami, a więc jego „zbędność" jako parametru wnoszącego nową porcję dyskryminacji do modelu), wysokie świadczą o tym, że parametr przyczynia się istotnie do polepszenia dyskryminacji między badanymi grupami. Współczynnik lambda Wilks'a, który oscyluje między wartościami 0 (*idealna dyskryminacja*) i 1 (*brak dyskryminacji*), wskazuje jak bardzo wyłączenie danego parametru wpływa na pogorszenie dyskryminacji między grupami.

Charakterystykę zmiennych włączonych do modelu podaje tabela:

Zestawienie analizy funkcji dyskryminacyjnej. Liczba zmiennych w modelu: 10; Zmienna grupująca: rodzaj zawału (3 grupy); Lambda Wilksa: 0.51055 przybl. F (20. 302) = 6.0328 p« 0.00001.

zmienna	lambda Wilksa	cząstkowe lambda Wilksa	$F_{usunięcia}$ (2.151)	poziom *p*	tolerancja	1-tolerancja (R^2)
BMI	0.535	**0.955**	3.5600	**0.031**	0.892	**0.108**
HDL	0.516	0.990	0.7950	0.453	0.828	0.172
LDL	0.760	**0.671**	36.9380	**8.73×10^{-14}**	0.899	**0.100**
TG	0.529	0.965	2.7360	0.068	0.726	0.274
Wiek	0.511	0.999	0.0120	0.988	0.862	0.138
PAI-1	0.609	**0.838**	14.5790	**1.63×10^{-6}**	0.921	**0.079**
Fg	0.511	0.999	0.0603	0.942	0.935	0.065
Plt	0.512	0.996	0.2780	0.757	0.939	0.061
ciśnienie skurczowe	0.511	0.999	0.0250	0.976	0.373	0.627
ciśnienie rozkurczowe	0.513	0.996	0.3220	0.725	0.405	0.595

Widzimy, że zmiennymi, które w sposób najbardziej znaczący przyczyniły się do dyskryminacji grup były: LDL (wartość statystyki F_{usuw} >36, p«0.0001), PAI-1 (F_{usuw} >14, p«0.0001) oraz BMI (F_{usuw} >3.5, p <0.05).

Niskie wartości współczynnika lambda cząstkowego dla LDL i PAI-1. a w mniejszym stopniu także BMI, są wskazaniem, że parametry te przyczyniają się istotnie do polepszenia dyskryminacji między grupami pacjentów. Parametry te odznaczają się wysoką tolerancją (czyli niewielką zbędnością, R^2 mniejsze lub w przybliżeniu równe 0.100). Wysoką tolerancją odznaczają się także stężenie fibrynogenu oraz miano płytek (R^2 <0.100). Ten pierwszy parametr znajduje się zresztą na granicy istotności, jeśli chodzi o dyskryminację grup (p = 0.06), wartości cząstkowe lambda Wilksa są jednak bliskie jedności, co wskazuje, że usunięcie tych parametrów nie pogorszyłoby istotnie dyskryminacji między grupami pacjentów. Najniższe wartości tolerancji zaobserwowano dla ciśnienia skurczowego i ciśnienia rozkurczowego, co wskazuje, że każdy z tych parametrów jest wysoce zbędny w modelu. Nie powinno to dziwić, ponieważ oba te parametry są oczywiście wysoce zależne od siebie (wysoce skorelowane).

Obliczamy następnie odległości Mahalanobisa, które charakteryzują istotność statystyczną dyskryminacji między porównywanymi grupami pacjentów.

Kwadraty odległości Mahalanobisa.

	bez zawału	**zawał niepełnościenny**	**zawał pełnościenny**
bez zawału		4.165	2.478
zawał niepełnościenny	4.165		1.245
zawał pełnościenny	2.478	1.245	

Wartości statystyki *F*; *df* = 10, 151.

	bez zawału	**zawał niepełnościenny**	**zawał pełnościenny**
bez zawału		9.036	7.270
zawał niepełnościenny	9.036		2.565
zawał pełnościenny	7.270	2.565	

Istotność dyskryminacji między grupami (poziom *p*).

	bez zawału	**zawał niepełnościenny**	**zawał pełnościenny**
bez zawału		0.000	0.000
zawał niepełnościenny	0.000		0.007
zawał pełnościenny	0.000	0.007	

Wartości kwadratowych odległości Mahalanobisa między punktami centralnymi (*centroidami*) trzech badanych grup pacjentów nie posiadają wysokich wartości bezwzględnych, ale – jak widzimy – wartości testu *F* oraz znamienność statystyczna dyskryminacji między pacjentami bez zawału, z zawałem niepełnościennym oraz z zawałem pełnościennym jest bardzo wysoka we wszystkich porównaniach międzygrupowych.

Następnym krokiem jest obliczenie wartości standaryzowanych współczynników regresji funkcji dyskryminacyjnej opisującej udział wszystkich włączonych do modelu parametrów w separacji (dyskryminacji) poszczególnych grup:

Współczynniki standaryzowane dla zmiennych kanonicznych.

	funkcja dyskryminacyjna 1	**funkcja dyskryminacyjna 2**
BMI	–0.21677	–0.49942
HDL	–0.15046	–0.15913
LDL	0.919431	0.317425
TG	0.317133	–0.22341
wiek	0.016356	–0.02529
PAI–1	0.431852	–0.88929
Fg	–0.04229	0.02942
Plt	–0.02422	–0.17205
ciśnienie skurczowe	0.029053	–0.06471
ciśnienie rozkurczowe	0.151307	0.086245
wartość własna	0.715976	0.141433
proporcja skumulowana	0.835047	1.000000

Wiadomo, że te zmienne, którym przypisane są największe wartości standaryzowanych współczynników regresji funkcji dyskryminacyjnej, przyczyniają się w najistotniejszym stopniu do poprawnego zaszeregowania badanego przypadku do jednej z badanych grup. W naszym przypadku są to LDL, PAI-1, BMI, a także Fg.

Na podstawie obliczonych współczynników funkcji dyskryminacyjnych możemy ponadto ustalić punkty centralne dla każdej z grup (*centroidy*). Pod względem rachunkowym punkty takie są uśrednionymi wartościami funkcji dyskryminacyjnej opisującej wszystkie zmienne w modelu w przestrzeni wielowymiarowej (*odległości Mahalanobisa*); można je sobie wyobrazić jako geometryczne środki elipsoid otaczających zbiorowości („chmury") punktów wyznaczonych przez wartości funkcji dyskryminacyjnej dla poszczególnych przypadków w modelu.

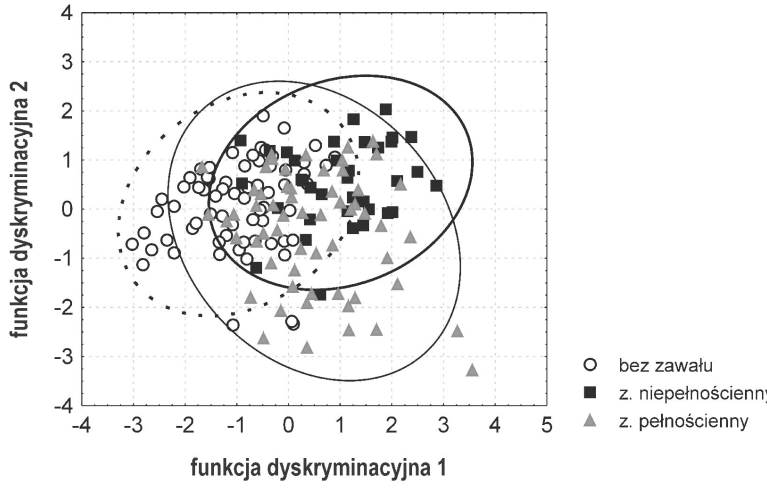
Średnie zmiennych kanonicznych.

	funkcja dyskryminacyjna 1	**funkcja dyskryminacyjna 2**
bez zawału	–0.96524	0.099548
zawał niepełnościenny	1.008673	0.537383
zawał pełnościenny	0.497017	–0.44263

Jeżeli w analizie dyskryminacyjnej uwzględniamy *n* grup badanych, to możliwe jest zbudowanie *n – 1* niezależnych (*ortogonalnych*) funkcji dyskryminujących (tzw. *funkcji kanonicznych lub pierwiastków*), które umożliwiają najlepszą możliwą dyskryminację pomiędzy daną grupą a każdą z grup pozostałych. O wiarygodności takiej dyskryminacji decydują istotności statystyczne funkcji dyskryminacyjnych.

Testy chi-kwadrat dla kolejnych pierwiastków.

	wartość własna	**współczynnik korelacji zmiennych kanonicznych**	**cząstkowe lambda Wilksa**	**chi^2**	***df***	**poziom *p***
funkcja 1	0.715976	0.645942	0.51055	104.5374	20.0000	2.04×10^{-13}
funkcja 2	0.141433	0.352006	0.876092	20.57017	9.00000	0.014724



Analiza funkcji dyskryminacyjnej to przede wszystkim wspaniałe narzędzie do klasyfikacji przypadków do różnych porównywanych grup z jak największą trafnością. Aby to

zrobić posługujemy się wartościami funkcji klasyfikacyjnych; jest ich tyle ile badanych grup w modelu.

Współczynniki dla funkcji klasyfikacyjnych; zmienna grupująca: rodzaj zawału

	bez zawału	**zawał niepełnościenny**	**zawał pełnościenny**
BMI	2.534627	2.353233	2.521668
HDL	0.464741	0.433988	0.453524
LDL	0.13776	0.192731	0.170744
TG	0.043388	0.048314	0.048842
Wiek	0.531315	0.533428	0.535064
PAI-1	0.141369	0.155819	0.176119
Fg	3.304864	3.248659	3.242929
Plt	0.067618	0.065426	0.068648
ciśnienie skurczowe	0.064445	0.065706	0.067818
ciśnienie rozkurczowe	0.486824	0.516436	0.502183
stała	–115.392	–122.667	–124.502

Funkcje te obliczane są na podstawie wartości poszczególnych zmiennych dla poszczególnych przypadków, toteż dla każdego – nawet wątpliwego – przypadku możemy także obliczyć wartość funkcji klasyfikacyjnej. To one ułatwiają nam decyzję o przynależności badanego przypadku do którejś z porównywanych grup w modelu. Funkcje klasyfikacyjne mogą być bezpośrednio wykorzystane do obliczenia wartości klasyfikacyjnych dla jakichkolwiek nowych obserwacji. Czym wyższa wartość funkcji klasyfikacyjnej dla określonej grupy, tym wyższe prawdopodobieństwo, że dany przypadek należy właśnie do tej grupy.

Opierając się na wartościach tych funkcji dla poszczególnych przypadków włączonych do modelu możemy obliczyć tzw. prawdopodobieństwa *a posteriori*, czyli szanse, że dany przypadek zostanie zaklasyfikowany właśnie do grupy, do której go włączyliśmy. Takie prawdopodobieństwo, że dany przypadek będzie należał do danej grupy jest z grubsza proporcjonalne do odległości Mahalanobisa od punktu centralnego (centroidy) grupy; nie jest to ścisła proporcjonalność, gdyż na wartość punktu centralnego (centroidy) ma wpływ wiele zmiennych, zatem rozkład normalny wokół każdego centroidy jest „zaburzony" przez wpływ wielu zmiennych w modelu. W rzeczywistości, położenie każdego przypadku obliczamy na podstawie naszej wcześniejszej wiedzy o wartościach, jakie wszystkie zmienne należące do modelu przyjmują dla danego przypadku, i tak szacowane prawdopodobieństwa określamy jako prawdopodobieństwa *a posteriori* w odróżnieniu od prawdopodobieństw *a priori*, ocenianych na podstawie naszych przypuszczeń o przynależności danego przypadku do określonej grupy.

Dla naszego modelu i uwzględnionych w nim zmiennych oraz przypadków obliczone prawdopodobieństwa *a posteriori* wynoszą:

Prawdopodobieństwa *a posteriori*. Błędne klasyfikacje są oznaczone *x*.

przypadek nr	**klasyfikacja *a priori***	**bez zawału** $p = 0.417$	**zawał niepełnościenny** $p = 0.221$	**zawał pełnościenny** $p = 0.362$
1	*bez zawału*	0.710	0.131	0.159
* **3**	*bez zawału*	0.311	0.374	0.316
4	*bez zawału*	0.868	0.027	0.106
5	*bez zawału*	0.931	0.004	0.066
7	*bez zawału*	0.829	0.071	0.100

przypadek nr	klasyfikacja *a priori*	bez zawału $p = 0.417$	zawał niepełnościenny $p = 0.221$	zawał pełnościenny $p = 0.362$
9	*bez zawału*	0.448	0.212	0.340
10	*bez zawału*	0.848	0.036	0.116
11	*bez zawału*	0.921	0.016	0.063
12	*bez zawału*	0.925	0.020	0.055
13	*bez zawału*	0.589	0.054	0.357
15	*bez zawału*	0.970	0.004	0.027
16	*bez zawału*	0.803	0.018	0.178
18	*bez zawału*	0.908	0.019	0.074
19	*bez zawału*	0.904	0.010	0.086
20	*bez zawału*	0.970	0.003	0.027
21	*bez zawału*	0.566	0.214	0.220
22	*bez zawału*	0.949	0.003	0.048
23	*bez zawału*	0.668	0.145	0.186
25	*bez zawału*	0.461	0.078	0.461
26	*bez zawału*	0.790	0.051	0.159
27	*bez zawału*	0.665	0.039	0.296
28	*bez zawału*	0.967	0.001	0.032
29	*bez zawału*	0.948	0.009	0.042
31	*bez zawału*	0.841	0.043	0.117
32	*bez zawału*	0.732	0.065	0.204
33	*bez zawału*	0.822	0.027	0.151
34	*bez zawału*	0.942	0.013	0.045
35	*bez zawału*	0.904	0.028	0.067
37	*bez zawału*	0.870	0.019	0.112
38	*bez zawału*	0.583	0.102	0.315
39	*bez zawału*	0.720	0.082	0.198
* **40**	*bez zawału*	0.232	0.499	0.269
41	*bez zawału*	0.975	0.001	0.024
42	*bez zawału*	0.570	0.134	0.296
43	*bez zawału*	0.780	0.017	0.203
45	*bez zawału*	0.905	0.024	0.071
* **46**	*bez zawału*	0.307	0.334	0.359
* **47**	*bez zawału*	0.382	0.175	0.443
49	*bez zawału*	0.608	0.035	0.358
* **50**	*bez zawału*	0.175	0.033	0.792
51	*bez zawału*	0.762	0.034	0.204
* **52**	*bez zawału*	0.283	0.317	0.400
* **54**	*bez zawału*	0.344	0.087	0.568
* **55**	*bez zawału*	0.187	0.034	0.779
56	*bez zawału*	0.621	0.192	0.187
* **57**	*bez zawału*	0.177	0.478	0.346
58	*bez zawału*	0.758	0.092	0.150
60	*bez zawału*	0.900	0.025	0.075
61	*bez zawału*	0.610	0.089	0.301
62	*bez zawału*	0.543	0.010	0.447
65	*bez zawału*	0.632	0.165	0.203
67	*bez zawału*	0.658	0.174	0.168
68	*bez zawału*	0.636	0.243	0.122
70	*bez zawału*	0.460	0.355	0.185
71	*bez zawału*	0.576	0.186	0.237
72	*bez zawału*	0.571	0.085	0.344
* **76**	*bez zawału*	0.376	0.104	0.520
78	*bez zawału*	0.449	0.253	0.298
79	*bez zawału*	0.909	0.009	0.082
80	*bez zawału*	0.964	0.002	0.035

przypadek nr	klasyfikacja *a priori*	bez zawału $p = 0.417$	zawał niepełnościenny $p = 0.221$	zawał pełnościenny $p = 0.362$
84	*bez zawału*	0.980	0.001	0.019
85	*bez zawału*	0.779	0.025	0.196
*** 87**	*bez zawału*	0.318	0.124	0.558
89	*bez zawału*	0.955	0.005	0.040
90	*bez zawału*	0.643	0.066	0.291
91	*bez zawału*	0.680	0.031	0.289
100	*bez zawału*	0.658	0.109	0.232
*** 101**	*bez zawału*	0.135	0.553	0.312
103	*zawał niepełnościenny*	0.009	0.744	0.247
*** 104**	*zawał niepełnościenny*	0.026	0.481	0.494
*** 105**	*zawał niepełnościenny*	0.379	0.327	0.294
106	*zawał niepełnościenny*	0.022	0.499	0.479
*** 107**	*zawał niepełnościenny*	0.047	0.446	0.508
*** 108**	*zawał niepełnościenny*	0.054	0.354	0.592
*** 109**	*zawał niepełnościenny*	0.192	0.344	0.464
110	*zawał niepełnościenny*	0.009	0.842	0.149
111	*zawał niepełnościenny*	0.128	0.548	0.324
*** 112**	*zawał niepełnościenny*	0.442	0.305	0.253
114	*zawał niepełnościenny*	0.005	0.729	0.266
*** 115**	*zawał niepełnościenny*	0.078	0.447	0.474
*** 116**	*zawał niepełnościenny*	0.238	0.155	0.607
117	*zawał niepełnościenny*	0.049	0.724	0.226
*** 118**	*zawał niepełnościenny*	0.322	0.305	0.373
*** 119**	*zawał niepełnościenny*	0.090	0.369	0.540
120	*zawał niepełnościenny*	0.018	0.658	0.323
121	*zawał niepełnościenny*	0.020	0.866	0.114
*** 123**	*zawał niepełnościenny*	0.072	0.317	0.612
*** 124**	*zawał niepełnościenny*	0.476	0.140	0.384
125	*zawał niepełnościenny*	0.032	0.742	0.226
126	*zawał niepełnościenny*	0.065	0.766	0.168
128	*zawał niepełnościenny*	0.019	0.794	0.187
*** 129**	*zawał niepełnościenny*	0.116	0.079	0.805
*** 131**	*zawał niepełnościenny*	0.056	0.466	0.478
132	*zawał niepełnościenny*	0.094	0.515	0.391
134	*zawał niepełnościenny*	0.018	0.807	0.174
135	*zawał niepełnościenny*	0.133	0.612	0.255
*** 137**	*zawał niepełnościenny*	0.237	0.219	0.544
*** 138**	*zawał niepełnościenny*	0.518	0.039	0.443
139	*zawał niepełnościenny*	0.088	0.556	0.356
*** 140**	*zawał niepełnościenny*	0.312	0.308	0.380
*** 141**	*zawał niepełnościenny*	0.793	0.102	0.105
*** 142**	*zawał niepełnościenny*	0.589	0.210	0.201
*** 151**	*zawał niepełnościenny*	0.760	0.071	0.169
*** 154**	*zawał niepełnościenny*	0.259	0.319	0.422
155	*zawał pełnościenny*	0.150	0.058	0.791
156	*zawał pełnościenny*	0.185	0.268	0.547
*** 157**	*zawał pełnościenny*	0.716	0.033	0.251
*** 158**	*zawał pełnościenny*	0.638	0.086	0.277
159	*zawał pełnościenny*	0.301	0.016	0.682
160	*zawał pełnościenny*	0.145	0.072	0.783
*** 161**	*zawał pełnościenny*	0.294	0.410	0.296
162	*zawał pełnościenny*	0.238	0.062	0.700
163	*zawał pełnościenny*	0.101	0.026	0.873
*** 164**	*zawał pełnościenny*	0.107	0.574	0.319
*** 165**	*zawał pełnościenny*	0.922	0.022	0.056
*** 166**	*zawał pełnościenny*	0.449	0.253	0.297

przypadek nr	klasyfikacja *a priori*	bez zawału $p = 0.417$	zawał niepełnościenny $p = 0.221$	zawał pełnościenny $p = 0.362$
167	*zawał pełnościenny*	0.259	0.034	0.706
169	*zawał pełnościenny*	0.031	0.405	0.564
* **170**	*zawał pełnościenny*	0.877	0.017	0.106
171	*zawał pełnościenny*	0.072	0.099	0.829
* **172**	*zawał pełnościenny*	0.016	0.654	0.330
* **173**	*zawał pełnościenny*	0.067	0.503	0.430
* **174**	*zawał pełnościenny*	0.571	0.214	0.215
* **179**	*zawał pełnościenny*	0.611	0.165	0.224
183	*zawał pełnościenny*	0.043	0.111	0.846
* **184**	*zawał pełnościenny*	0.393	0.232	0.375
185	*zawał pełnościenny*	0.053	0.414	0.533
186	*zawał pełnościenny*	0.442	0.105	0.453
187	*zawał pełnościenny*	0.002	0.157	0.842
188	*zawał pełnościenny*	0.012	0.428	0.560
189	*zawał pełnościenny*	0.254	0.083	0.663
190	*zawał pełnościenny*	0.001	0.090	0.909
191	*zawał pełnościenny*	0.422	0.059	0.519
193	*zawał pełnościenny*	0.114	0.384	0.502
194	*zawał pełnościenny*	0.171	0.147	0.682
195	*zawał pełnościenny*	0.023	0.279	0.698
196	*zawał pełnościenny*	0.146	0.397	0.457
197	*zawał pełnościenny*	0.383	0.212	0.406
* **198**	*zawał pełnościenny*	0.034	0.716	0.250
* **199**	*zawał pełnościenny*	0.540	0.074	0.386
200	*zawał pełnościenny*	0.421	0.145	0.434
* **201**	*zawał pełnościenny*	0.037	0.752	0.212
* **202**	*zawał pełnościenny*	0.425	0.218	0.357
* **203**	*zawał pełnościenny*	0.571	0.057	0.373
* **204**	*zawał pełnościenny*	0.796	0.030	0.174
205	*zawał pełnościenny*	0.081	0.390	0.529
* **206**	*zawał pełnościenny*	0.565	0.209	0.226
207	*zawał pełnościenny*	0.118	0.204	0.678
* **208**	*zawał pełnościenny*	0.087	0.649	0.264
210	*zawał pełnościenny*	0.486	0.023	0.490
* **211**	*zawał pełnościenny*	0.624	0.110	0.267
212	*zawał pełnościenny*	0.072	0.426	0.501
* **213**	*zawał pełnościenny*	0.186	0.447	0.367
214	*zawał pełnościenny*	0.014	0.205	0.781
* **215**	*zawał pełnościenny*	0.514	0.131	0.355
* **216**	*zawał pełnościenny*	0.107	0.533	0.360
218	*zawał pełnościenny*	0.049	0.090	0.862
* **222**	*zawał pełnościenny*	0.767	0.040	0.193
223	*zawał pełnościenny*	0.039	0.058	0.903
224	*zawał pełnościenny*	0.254	0.127	0.619
* **225**	*zawał pełnościenny*	0.673	0.096	0.231
226	*zawał pełnościenny*	0.018	0.078	0.904
237	*zawał pełnościenny*	0.292	0.210	0.498

Widzimy, że pojedyncze przypadki „nie trafienia" spotykamy w grupie bez zawału, natomiast bardzo wiele błędnie sklasyfikowanych obserwacji występuje w grupie z zawałem niepełnościennym i zawałem pełnościennym.

Naszym podsumowującym sprawdzianem jak dobra jest predykcja w oparciu o zbudowany przez nas model, to znaczy na ile dobrze obliczone funkcje klasyfikacyjne pozwalają przewidzieć przynależność przypadków do określonej grupy, obliczamy macierz klasyfika-

cji. Pokazuje ona liczbę przypadków, które zostały poprawnie zaklasyfikowane do określonej grupy oraz tych, które zostały błędnie zaklasyfikowane.

Macierz klasyfikacji. Wiersze: obserwowana klasyfikacja; kolumny: przewidywana klasyfikacja.

	jaki procent poprawnie sklasyfikowanych	**bez zawału** $p = 0.417$	**zawał niepełnościenny** $p = 0.221$	**zawał pełnościenny** $p = 0.362$
bez zawału	82.4	56.0	4.0	8.0
zawał niepełnościenny	41.7	7.0	15.0	14.0
zawał pełnościenny	55.9	17.0	9.0	33.0
razem	**63.8**	**80.0**	**28.0**	**55.0**

W pierwszej kolumnie tabeli (opisanej *„jaki procent poprawnie sklasyfikowanych"*) podano trafność naszej klasyfikacji *a priori*. Na przykład, do grupy 1 (bez zawału) zaklasyfikowaliśmy *a priori* 68 przypadków, z czego – na podstawie szacowania w oparciu o wartości funkcji klasyfikacyjnych – pozostało w tej grupie 56 przypadków, zatem poprawność klasyfikacji wynosi: 56/68 = 0.824. Podobnie, do grupy 2 zaklasyfikowaliśmy wstępnie 36 przypadków, z których jedynie 15 okazało się być „poprawnie" (*a posteriori*, w oparciu o wartości funkcji klasyfikacyjnych) przypisanych, czyli trafność naszej klasyfikacji wynosi 15/36 = 0.4166. itd. Pod nagłówkiem z nazwami grup pacjentów zamieszczono z kolei prawdopodobieństwa klasyfikacji przewidywanej. Na przykład, w grupie bez zawału mamy (56 + 4 + 8) = 68 przypadków wśród 163 wszystkich rozpatrywanych przypadków, co daje 68/163 = 0.417. Analogicznie, w grupie z zawałem pełnościennym mamy (17 + 9 + 33) = 59 przypadków, czyli 59/163 = 0.3619 całości, itd.

Widzimy, że najbardziej poprawnie zostali *a priori* sklasyfikowani pacjenci bez zawału, najbardziej błędnie ci z zawałem niepełnościennym. Jedynie około 64% wszystkich przypadków zostało poprawnie sklasyfikowanych *a priori*. Zauważmy też, że ponieważ wybrano opcję usuwana przypadków z brakującymi danymi, analiza objęła 163 ważne przypadki. Macierz klasyfikacji mówi nam także wiele o jakości dyskryminacji między grupami. W naszym przypadku elipsy rozrzutu wokół punktów wyznaczonych przez wartości funkcji dyskryminacyjnych zachodzą znacznie na siebie, pomimo istotności dyskryminacji między grupami. Te obszary nakładania się elips rozrzutu stanowią właśnie o naszym braku precyzji w klasyfikacji przypadków *a priori*. Gdyby elipsy rozrzutu były znacznie oddalone od siebie (nie występowałaby interferencja między grupami pacjentów), to macierz klasyfikacji ujawniłaby znacznie wyższą trafność klasyfikacji przypadków do grup. Czym lepiej oddzielona zbiorowość punktów funkcji dyskryminacyjnej, a tym samym i elips rozrzutu, tym wyższy procent poprawnie sklasyfikowanych przypadków.

W naszym przykładzie zastosowaliśmy standardowy model analizy, który analizował udział wszystkich parametrów (i tych przyczyniających się istotnie do dyskryminacji między grupami i tych nie poprawiających dyskryminacji). Możemy jednak od razu na wstępie dokonać selekcji zmiennych oraz uwzględnić jedynie te, które istotnie poprawiają separację grup. Służą do tego modele krokowe analizy funkcji dyskryminacyjnej (postępującej lub wstecznej). Pierwszy włącza kolejne zmienne do modelu do czasu, aż nowa zmienna nie poprawia już istotnie dyskryminacji między grupami. Drugi włącza na wstępie wszystkie zmienne do modelu, a następnie eliminuje kolejne zmienne, których usunięcie poprawia dopasowanie modelu.

Zastosujmy oba modele do naszych danych, aby przekonać się czy wyniki krokowej analizy będą spójne z wynikami uzyskanymi przy zastosowaniu modelu standardowego. Dla modelu krokowego postępującego mamy:

Zestawienie analizy funkcji dyskryminacyjnej krokowej postępującej. Liczba zmiennych w modelu: 4; zmienna grupująca: rodzaj zawału (3 grupy); Lambda Wilksa: 0.52506 przybl. F (8, 314) = 14.917 p< 0.00001

	lambda Wilksa	cząstkowe lambda Wilksa	$F_{\text{usunięcia}}$ (2, 157)	poziom p	tolerancja	1-tolerancja (R^2)
LDL	0.791	0.664	39.692	0.000	0.919	0.081
PAI-1	0.625	0.840	14.959	0.000	0.953	0.047
TG	0.556	0.944	4.657	0.011	0.875	0.125
BMI	0.548	0.959	3.361	0.037	0.931	0.069

Do modelu zostały włączone tylko te zmienne, które istotnie przyczyniają się do dyskryminacji między grupami pacjentów. Najniższe wartości cząstkowego lambda Wilksa stwierdzono dla LDL i PAI-1 – to te dwie zmienne posiadają największą moc dyskryminacyjną. W modelu nie znalazły się pozostałe zmienne, które charakteryzują się niskimi wartościami statystyki F (i w związku z tym niską istotnością) oraz wysokimi wartościami statystyki cząstkowego lambda Wilksa:

Zmienne aktualnie poza modelem.

	lambda Wilksa	cząstkowe lambda Wilksa	$F_{\text{wprowadzenia}}$ (2, 156)	poziom p	tolerancja	1-tolerancja (R^2)
HDL	0.520	0.990	0.785	0.458	0.850	0.150
wiek	0.525	0.999	0.056	0.946	0.947	0.053
Fg	0.525	0.999	0.072	0.930	0.964	0.036
Plt	0.522	0.995	0.395	0.674	0.990	0.010
WBC	0.523	0.996	0.294	0.746	0.932	0.068
ciśnienie skurczowe	0.521	0.992	0.631	0.534	0.960	0.040
cisnienie rozkurczowe	0.519	0.989	0.891	0.412	0.974	0.026

Czy taki model zapewnia lepszą dyskryminację grup?

Aby się przekonać liczymy odległości Mahalanobisa oraz istotność tych odległości między centroidami.

Kwadraty odległości Mahalanobisa.

	bez zawału	zawał niepełnościenny	zawał pełnościenny
bez zawału		3.905	2.402
zawał niepełnościenny	3.905		1.145
zawał pełnościenny	2.402	1.145	

Istotność dyskryminacji między grupami (poziom istotności p).

	bez zawału	zawał niepełnościenny	zawał pełnościenny
bez zawału		0.000	0.000
zawał niepełnościenny	0.000		0.000
zawał pełnościenny	0.000	0.000	

W tym modelu istotność różnic między grupami jest wyższa, szczególnie między grupą pacjentów z zawałem niepełnościennym oraz pacjentów z zawałem pełnościennym. Pomimo tego, trafność klasyfikacji *a priori* nie uległa poprawie:

Macierz klasyfikacji. Wiersze: obserwowana klasyfikacja; kolumny: przewidywana klasyfikacja.

	jaki procent poprawnie sklasyfikowanych	**bez zawału** $p = 0.417$	**zawał niepełnościenny** $p = 0.221$	**zawał pełnościenny** $p = 0.362$
bez zawału	83.1	69.0	6.0	8.0
zawał niepełnościenny	39.6	11.0	19.0	18.0
zawał pełnościenny	51.3	22.0	17.0	41.0
razem	**61.1**	**102.0**	**42.0**	**67.0**

Dla modelu krokowego wstecznego uzyskaliśmy:

Zestawienie analizy funkcji dyskryminacyjnej krokowej wstecznej. Liczba zmiennych w modelu: 2. po 9 krokach; zmienna grupująca: rodzaj zawału (3 grupy); Lambda Wilksa: 0.57780 przybl. F (4.318) = 25.088 p< 0.00001.

	lambda Wilksa	**cząstkowe lambda Wilksa**	$F_{\text{usunięcia}}$ **(2, 159)**	**poziom *p***	**tolerancja**	**1-tolerancja (R^2)**
LDL	0.831	0.695	34.865	0.000	1.000	0.000
PAI-1	0.675	0.856	13.409	0.000	1.000	0.000

Widzimy, że ten algorytm analizy dyskryminacyjnej uwzględnił jedynie udział dwóch zmiennych w modelu – tych, które w sposób najistotniejszy przyczyniają się do separacji trzech grup pacjentów. Pozostałe zmienne znalazły się poza modelem:

Zmienne aktualnie poza modelem.

	lambda Wilksa	**cząstkowe lambda Wilksa**	$F_{\text{wprowadzenia}}$ **(2, 157)**	**poziom *p***	**tolerancja**	**1-tolerancja (R^2)**
wiek	0.577	0.999	0.043	0.958	0.994	0.006
Fg	0.576	0.997	0.206	0.814	0.992	0.008
ciśnienie skurczowe	0.571	0.989	0.899	0.409	0.964	0.036
WBC	0.571	0.989	0.919	0.401	0.968	0.032
Plt	0.575	0.995	0.389	0.678	0.991	0.009
HDL	0.566	0.980	1.604	0.204	0.999	0.001
ciśnienie rozkurczowe	0.570	0.987	1.041	0.356	0.975	0.025
BMI	0.556	0.963	3.066	0.049	0.966	0.034
TG	0.548	0.948	4.365	0.014	0.907	0.093

Widzimy, że wśród nich są także i te, które w sposób istotny przyczyniają się do dyskryminacji (BMI, TG). Obie te zmienne mają jednak wysokie wartości współczynnika cząstkowego lambda Wilksa, co sugeruje, że ich usunięcie z modelu nie pogarsza istotnie dyskryminacji między grupami. Możemy to ocenić spoglądając na istotność odległości Mahalanobisa między centroidami:

Istotność dyskryminacji między grupami (poziom istotności p).

	bez zawału	**zawał niepełnościenny**	**zawał pełnościenny**
bez zawału		0.000	0.000
zawał niepełnościenny	0.000		0.000
zawał pełnościenny	0.000	0.000	

Nie poprawia się też istotnie trafność klasyfikacji przypadków – nieznacznie lepszą predykcję zauważamy jedynie dla grupy 1 – bez zawału:

Macierz klasyfikacji. Wiersze: obserwowana klasyfikacja; kolumny: przewidywana klasyfikacja.

	jaki procent poprawnie sklasyfikowanych	**bez zawału** $p = 0.417$	**zawał niepełnościenny** $p = 0.221$	**zawał pełnościenny** $p = 0.362$
bez zawału	84.9	73.0	6.0	7.0
zawał niepełnościenny	36.7	11.0	18.0	20.0
zawał pełnościenny	48.8	25.0	16.0	39.0
razem	**60.5**	**109.0**	**40.0**	**66.0**

Możemy zatem wnioskować, że dla badanej zbiorowości wyników w trzech grupach pacjentów, niezależnie od dobrania metody analizy funkcji dyskryminacyjnej, ostateczne wyniki takiej analizy są bardzo podobne.

Regresja logistyczna

Przykład 91

Na podstawie danych zebranych wśród 260 pacjentów z chorobą niedokrwienną serca (wyniki w arkuszu Excela ***regresja logistyczna.xls***) badano wpływ wybranych czynników na ryzyko wystąpienia zawału mięśnia sercowego. Do analizy włączono następujące zmienne:

zawał wystąpienie zawału (1 – zawał, 0 – bez zawału), cecha dichotomiczna (dyskretna), zmienna zależna jakościowa,

HPA-1 polimorfizm glikoproteiny IIIa płytek krwi (1 – Pl^{A2}(+), 0 – Pl^{A2}(–)), cecha dichotomiczna (dyskretna), zmienna niezależna jakościowa,

dyslipidemia występowanie dyslipidemii (1 – jest, 0 – brak), cecha dichotomiczna (dyskretna), zmienna niezależna jakościowa,

palenie 1 – pali, 0 – nie pali, cecha dichotomiczna (dyskretna),

choroba wieńcowa występowanie choroby wieńcowej (1 – jest, 0 – nie ma, cecha dichotomiczna (dyskretna), zmienna niezależna jakościowa,

czas choroby wieńcowej czas trwania choroby wieńcowej (1 – <2 miesiące, 2 – 2-12 miesięcy, 3 – powyżej 12 miesięcy), cecha dyskretna, zmienna niezależna jakościowa,

BMI wskaźnik masy ciała, zmienna niezależna ilościowa,

wiek wiek pacjenta, zmienna niezależna ilościowa.

Analizę przeprowadzono z wykorzystaniem pakietu statystycznego Statistica v. 5.3 (*Statsoft*). Estymację równania regresji logistycznej wykonano metodą quasi-Newtona. Do liczenia funkcji straty zastosowano metodę największej wiarygodności.

Parametry tak estymowanego modelu podaje tabela:

Model: regresja logistyczna (logit): liczba 0:62 1: 115. Zmienna zależna: ZAWAŁ, strata: najw. wiaryg., błąd średniokw. skalowany do 1; końcowa strata 110.13; $chi^2(7) = 8.996$ $p = 0.25296$

	Stała B_0	**HPA-1**	**dyslipidemia**	**palenie**	**BMI**	**wiek**	**choroba wieńcowa**	**czas choroby wieńcowej**
ocena	–1.975	0.097	0.116	0.776	–0.014	0.013	0.427	0.459
błąd std.	2.047	0.352	0.341	0.353	0.046	0.017	1.446	0.248
t Studenta(169)	–0.965	0.275	0.340	2.202	–0.292	0.759	0.295	1.853
poziom *p*	0.336	0.784	0.734	**0.029**	0.770	0.449	0.768	**0.066**
–95%CL	–6.016	–0.598	–0.557	0.080	–0.105	–0.021	–2.429	–0.030
+95%CL	2.066	0.791	0.789	1.472	0.078	0.047	3.282	0.948
chi^2 Walda	0.931	0.075	0.116	4.848	0.085	0.576	0.087	3.434
poziom *p*	0.335	0.784	0.734	**0.028**	0.770	0.448	0.768	**0.064**
iloraz szans zmiany jednostkowej	0.139	1.101	1.123	2.173	0.987	1.013	1.532	1.583
–95%CL	0.002	0.550	0.573	1.084	0.900	0.979	0.088	0.970
+95%CL	7.897	2.205	2.201	4.359	1.081	1.049	26.632	2.581
iloraz szans zakresu		1.101	1.123	4.723	0.767	1.833	1.532	2.505
–95%CL		0.550	0.573	1.174	0.128	0.379	0.088	0.942
+95%CL		2.205	2.201	19.001	4.596	8.864	26.632	6.661

Analiza objęła 167 ważnych przypadków, z których zawał wystąpił w 115 przypadkach (64.97%) i nie wystąpił w 62 przypadkach (35.03%). Wartości logarytmu wiarygodności (-2log[wiarygodność modelu] jako miara dopasowania modelu) wynoszą 220.26 dla modelu z włączonymi zmiennymi oraz 229.26 dla modelu tylko z wyrazem wolnym. Statystyka chi^2, opisująca różnicę –2log[wiarygodność modelu ze zmiennymi] – (–2log[wiarygodność modelu tylko ze stałą]) dla 7 stopni swobody, okazała się nieistotna $p = 0.253$. Oznacza to, że włączenie do modelu takiego zestawu zmiennych nie poprawia dopasowania modelu, czyli nie wnosi nowej informacji do modelu.

W wierszu „ocena" znajdują się wartości współczynników regresji dla każdej zmiennej niezależnej. Widzimy, że współczynniki te posiadają najwyższe wartości dla zmiennych: palenie, czas choroby wieńcowej oraz choroba wieńcowa. Jedynie w przypadku palenia współczynnik kierunkowy jest istotny statystycznie. Znajduje to odzwierciedlenie w wartościach statystyki testu chi^2 Walda (jedynie palenie jest zmienną istotnie objaśniającą wariancję zmiennej zależnej, tzn. tłumaczącą wiarygodnie wystąpienie zawału), jak również w przedziałach ufności dla ilorazu szans. Iloraz ten jest wyrażony dwojako: jako wzrost ryzyka zawału przy jednostkowej zmianie zmiennej niezależnej (pali-nie pali) oraz jako wzrost ryzyka zawału przy zmianie zmiennej o wartość zakresu tej zmiennej. Oczywiście, dla zmiennych dichotomicznych te dwie wartości są identyczne, gdyż zmiana jednostkowa, to cały zakres zmienności takiej zmiennej. Różnica pojawia się dla zmiennych ciągłych lub zmiennych dyskretnych z liczbą kategorii większą niż 2. Na przykład, dla zmiennej „czas choroby wieńcowej" OR dla zmiany jednostkowej (o jedną kategorię, np. między grupą z chorobą wieńcową <2 miesięcy a grupą z chorobą wieńcową od 2 do 12 miesięcy) wynosi 1.583, natomiast OR dla zmiany o zakres (tzn. między grupą z chorobą wieńcową <2 miesięcy a grupą z chorobą wieńcową powyżej 12 miesięcy, różnica 2 kategorii) wynosi 2.505.

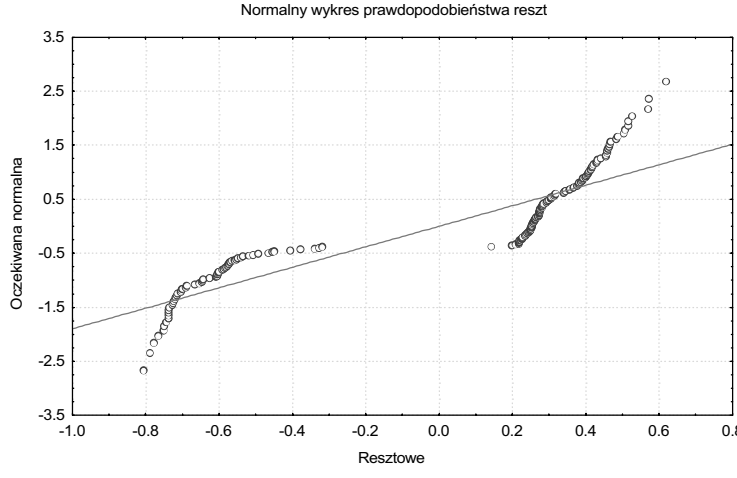
Zestawienie poprawnych i niepoprawnych klasyfikacji (wartości przewidywanych względem obserwowanych) zawiera tabela:

Klasyfikacja przypadków. Iloraz szans: 2.2712

obserwowane	przewidywane 0	przewidywane 1	procent poprawnych
0	9	53	14.5
1	8	107	93.0

Podczas gdy trafność klasyfikacji przypadków zawału była bardzo duża, to klasyfikacja pacjentów bez zawału była niepokojąco niska.

Wykres normalny prawdopodobieństwa dla reszt (Ryc. 21) pokazuje, że rozkład reszt silnie odbiega od linii prostej, czyli możemy wnioskować, że reszty nie mają rozkładu normalnego. Odbiegający od normalnego rozkład reszt wskazuje, że model obejmujący wybrane zmienne niezależne nie charakteryzuje się dobrą predykcją zmiennej zależnej (zawału) w całym zakresie przedziału ufności.



Ryc. 21. Normalny wykres prawdopodobieństwa reszt dla modelu regresji logistycznej. Zmienna zależna: zawał, zmienne niezależne: HPA1-2. dyslipidemia, palenie, choroba wieńcowa, czas choroby wieńcowej, BMI, wiek.

Wykonajmy teraz analizę dla tego samego zbioru danych, ale z uwzględnieniem jedynie niektórych zmiennych:

zawał wystąpienie zawału (1 – zawał, 0 – bez zawału), cecha dichotomiczna (dyskretna), zmienna zależna jakościowa,

HPA-1 polimorfizm glikoproteiny IIIa płytek krwi (1 – Pl^{A2}(+), 0 – Pl^{A2}(–)), cecha dichotomiczna (dyskretna), zmienna niezależna jakościowa,

palenie 1 – pali, 0 – nie pali, cecha dichotomiczna (dyskretna),

czas choroby wieńcowej czas trwania choroby wieńcowej (1 – <2 miesiące, 2 – 2-12 miesięcy, 3 – powyżej 12 miesięcy), cecha dyskretna, zmienna niezależna jakościowa,

wiek wiek pacjenta, zmienna niezależna ilościowa.

Parametry tak estymowanego modelu podaje tabela:

Model: regresja logistyczna (logit): liczba 0:66 1: 121. Zmienna zależna: ZAWAŁ, strata: najw. wiaryg., błąd średniokw. skalowany do 1; końcowa strata 116.43; $chi^2(4) = 9.96$ $p = 0.0412$.

	stała B_0	**HPA-1**	**palenie**	**wiek**	**czas choroby wieńcowej**
ocena	–1.827	0.117	0.779	0.011	0.477
błąd std.	1.109	0.346	0.345	0.017	0.240
t Studenta (169)	–1.647	0.337	2.258	0.674	1.988
poziom *p*	0.101	0.737	**0.025**	0.501	**0.048**
–95%CL	–4.015	–0.567	0.098	–0.022	0.004
+95%CL	0.361	0.800	1.460	0.044	0.951
chi^2 Walda	2.714	0.113	5.100	0.455	3.951
poziom *p*	0.099	0.736	**0.024**	0.500	**0.047**
iloraz szans zmiany jednostkowej	0.161	1.124	**2.179**	1.011	**1.612**
–95%CL	0.018	0.567	1.103	0.979	1.004
+95%CL	1.435	2.226	4.304	1.045	2.588
iloraz szans zakresu		1.124	2.179	1.678	2.597
–95%CL		0.567	1.103	0.369	1.007
+95%CL		2.226	4.304	7.632	6.698

Dla tego modelu statystyka chi^2 opisująca różnicę –2log[wiarygodność modelu ze zmiennymi] – (–2log[wiarygodność modelu tylko ze stałą]) dla 4 stopni swobody okazała się istotna, co oznacza, że włączenie do modelu takiego zestawu zmiennych istotnie poprawia jego dopasowanie i wnosi nową informację do modelu na temat czynników ryzyka wystąpienia zawału. Najwyższe wartości współczynników regresji charakteryzują zmienne: palenie oraz czas choroby wieńcowej. Widzimy więc, że znaczenie tych parametrów pozostało takie samo, jak w poprzednim modelu. Oba te współczynniki są istotne statystycznie, co oznacza wiarygodną predykcję wartości zmiennej zależnej (zawał) na podstawie wartości obu tych zmiennych niezależnych. Odpowiednio, wartości statystyki testu chi^2 Walda są także istotne dla obu zmiennych, czyli zarówno palenie, jak i czas choroby wieńcowej istotnie objaśniają wariancję zmiennej zależnej, tzn. wiarygodnie tłumaczą wystąpienie zawału. Iloraz szans wskazuje, że palenie zwiększa ryzyko zawału średnio 2.18 razy (od 1.1 do 4.3 razy z prawdopodobieństwem 95%), zaś długość czasu trwania choroby wieńcowej średnio 2.6 razy (od 1.0 do 6.7 razy, 95%).

Zestawienie poprawnych i niepoprawnych klasyfikacji (wartości przewidywanych względem obserwowanych) zawiera tabela:

Klasyfikacja przypadków. Iloraz szans: 2.489.

obserwowane	**przewidywane 0**	**przewidywane 1**	**procent poprawnych**
0	11	55	16.7
1	9	112	92.6

Iloraz szans obliczany jako stosunek iloczynu poprawnie sklasyfikowanych przypadków do iloczynu przypadków niepoprawnie sklasyfikowanych (112x11)/(9x55) = 2.489 jest wyższy od 1 i nieco wyższy niż dla poprzedniego modelu, co oznacza, że ta klasyfikacja jest lepsza, niż można byłoby oczekiwać przez czysty przypadek, oraz jest poprawniejsza niż dla poprzedniego modelu uwzględniającego więcej zmiennych.

Obliczmy teraz, o ile wzrośnie ryzyko zawału pod wpływem palenia papierosów oraz w miarę upływania czasu trwania choroby wieńcowej.

Iloraz szans dla dwóch rozważanych zmiennych (palenie papierosów: 0-1 oraz czas trwania choroby wieńcowej: 1-2-3) wyrazimy wzorem:

$$OR_{A\times B} = e^{b_1*(A-B)+b_2*(A-B)}$$

gdzie A oznacza wartość referencyjną, zaś B wartość porównywaną.

Cecha „palenie tytoniu" zmienia się skokowo w zakresie od 0 do 1 co jeden, zaś cecha: „czas trwania choroby wieńcowej" zmienia się także skokowo co jeden w zakresie od 1 do 3. Porównując pacjentów niepalących ze świeżą chorobą wieńcową (trwającą poniżej 2 miesięcy) z pacjentami, którzy palą i u których choroba trwa więcej niż 2 miesiące ale mniej niż rok, będziemy mieli:

$$OR_{A\times B} = e^{0.779*(1-0)+0.477*(2-1)} = e^{0.779+0.477} = e^{1.256} = 3.511$$

Ryzyko zawału wzrośnie u takich pacjentów ponad 3.5-krotnie. Gdybyśmy takie samo porównanie przeprowadzili po kolejnym roku (czyli w miarę jak czas choroby wieńcowej będzie się przedłużał), to uzyskalibyśmy:

$$OR_{A\times B} = e^{0.779*(1-0)+0.477*(3-1)} = e^{0.779+0.954} = e^{1.733} = 5.658$$

W takiej sytuacji ryzyko wzrosłoby ponad 5-krotnie.

Analiza log-liniowa

Przykład 92

Wśród pacjentów z chorobą niedokrwienną serca badano wpływ genetycznych polimorfizmów wybranych glikoprotein płytek krwi, płci oraz wieku na wystąpienie zawału mięśnia sercowego.

Celem badania było znalezienie czynników, które najsilniej wpływają na wystąpienie zawału, a przede wszystkim określenie, czy występowanie polimorfizmów VNTR(+) oraz HP^{A2}(+) glikoproteiny płytkowej Ib ma wpływ na częstość występowania zawału mięśnia sercowego.

Badano wpływ:

(4) polimorfizmu VNTR:	0 – VNTR-B (–), 1 – VNTR-B (+)
(5) polimorfizmu HPA-2:	0 – HPA-2b (–), 1 – HPA-2b (+),
(6) polimorfizmu ^{807}C/T:	0 – ^{807}T (–), 1 – ^{807}T (+); polimorfizm glikoproteiny Ia receptora dla kolagenu płytek krwi
(3) płci:	0 – kobiety, 1 – mężczyźni,
(2) kategorii wieku:	0 – poniżej 60 lat, 1 – powyżej 60 lat

na występowanie zawału (1).

Występowanie zawału było zmienną zależną od występowania kilku zmiennych objaśniających (zmiennych niezależnych).

Badaniami objęto 215 pacjentów (wyniki w arkuszu ***1analiza log-liniowa.xls***). W obliczeniach posłużymy się pakietem statystycznym Statistica PL v. 5.3 (*Statsoft*).

Na wstępie, w celu wybrania modelu i wyszczególnienia potencjalnych interakcji miedzy zmiennymi przeprowadzimy jednoczesne testy, że wszystkie interakcje między k czynnikami są jednocześnie równe 0. Wyniki tych testów przedstawia tabela:

Wyniki dotyczące wszystkich potencjalnych interakcji k czynników. Jednoczesne testy, że wszystkie interakcje k czynników są jednocześnie równe 0.

interakcje stopnia	**stopnie swobody**	**chi² najw. wiarygodn.**	**istotność *p***	**chi² Pearsona**	**istotność *p***
1	6	264.2319	0.0000	444.8501	0
2	15	208.9156	4.39×10^{-36}	230.7259	1.64×10^{-40}
3	20	27.49269	0.122068	40.78441	0.003987
4	15	4.220881	0.996917	1.866469	0.999981
5	6	2.705739	0.844762	1.005822	0.985389
6	1	7.21×10^{-5}	0.993226	7.35×10^{-5}	0.993159

Modele bez interakcji dwuczynnikowych ($p = 1.64 \times 10^{-40}$) oraz trójczynnikowych ($p = 0.003987$) należy odrzucić. Dołączenie interakcji rzędu 4 i wyższego nie daje istotnej poprawy dopasowania modelu ($p = 0.999981$). Wskazuje to, że model najmniej złożony to model bez interakcji czterowymiarowych i wyższych, natomiast uwzględniający interakcje rzędu drugiego i trzeciego, dlatego też dołączamy je do modelu. Wyniki w tabeli zależności brzegowych i zależności cząstkowych wskazują nam, które z nich są istotne:

Testy zależności brzegowych i cząstkowych.

	stopnie swobody	**zal. cząstkowe chi²**	**zal. cząstkowe *p***	**zal. brzegowe chi²**	**zal. brzegowe *p***
1	1	2.913544	0.087848	2.913544	0.087848
2	1	21.23090	4.09×10^{-6}	21.23090	4.09×10^{-6}
3	1	47.35156	6.04×10^{-12}	47.35156	6.04×10^{-12}
4	1	80.13651	3.68×10^{-19}	80.13651	3.68×10^{-19}
5	1	82.88031	9.21×10^{-20}	82.88031	9.21×10^{-20}
6	1	29.71924	5.03×10^{-8}	29.71924	5.03×10^{-8}
12	1	1.586933	0.207774	0.241364	0.623226
13	1	15.59652	7.86×10^{-5}	14.69489	0.000127
14	1	0.503040	0.478172	0.036133	0.849243
15	1	0.432346	0.510845	0.117462	0.731806
16	1	0.056267	0.812498	0.813446	0.367110
23	1	10.157590	0.001438	6.575775	0.010342
24	1	2.834816	0.092251	0.167267	0.682555
25	1	2.154320	0.142179	0.005096	0.943088
26	1	4.720200	0.029818	2.937195	0.086571
34	1	2.168541	0.14087	2.366333	0.123988
35	1	0.727882	0.393577	1.640228	0.200303
36	1	4.636539	0.031305	3.250916	0.071393
45	1	172.1910	3.1×10^{-39}	171.4982	4.39×10^{-39}
46	1	1.181377	0.277084	0.563538	0.452844
56	1	0.558605	0.454828	0.259796	0.610264
123	1	1.403995	0.236065	1.986938	0.158671
124	1	5.44×10^{-5}	0.994117	0.000549	0.981301
125	1	–0.001630	1.000000	0.206970	0.649156
126	1	0.373326	0.541201	0.334351	0.563113
134	1	–0.00243	1.000000	0.309235	0.578154
135	1	–0.00130	1.000000	0.821991	0.364606
136	1	0.000544	0.981399	0.020203	0.886974
145	1	4.265030	0.038913	5.982330	0.014455

146	1	–0.05116	1.000000	3.565796	0.058990
156	1	0.000483	0.982466	2.925415	0.087205
234	1	0.237482	0.626034	0.000244	0.987534
235	1	0.082686	0.773691	0.003387	0.953588
236	1	1.434885	0.230978	1.941559	0.163509
245	1	1.786064	0.181416	2.708542	0.099822
246	1	2.217061	0.136503	0.286652	0.592378
256	1	3.927474	0.047512	1.418640	0.233636
345	1	1.748879	0.186027	1.705040	0.191640
346	1	0.046842	0.828654	1.146729	0.284243
356	1	0.823389	0.364198	2.468811	0.116136
456	1	–0.07336	1.000000	1.342026	0.246686
1234	1	0.007662	0.930248	0.021790	0.882649
1235	1	0.003459	0.953098	0.032196	0.857599
1236	1	0.914190	0.339012	0.129425	0.719031
1245	1	–0.004970	1.000000	1.53×10^{-5}	0.996883
1246	1	–0.005130	1.000000	-6.1×10^{-5}	1.000000
1256	1	–0.008030	1.000000	0.198029	0.656320
1345	1	0.005684	0.939902	6.1×10^{-5}	0.993767
1346	1	0.003204	0.95486	0.573364	0.448931
1356	1	–0.00620	1.000000	0.041565	0.838453
1456	1	0.001760	0.966539	3.05×10^{-5}	0.995592
2345	1	–0.00037	1.000000	0.527168	0.467805
2346	1	0.004270	0.947901	1.497406	0.221080
2356	1	0.001901	0.965224	0.811310	0.367740
2456	1	0.778005	0.377759	1.049019	0.305741
3456	1	0.629543	0.427529	2.336266	0.126402
12345	1	-6.9×10^{-5}	1.000000	2.67×10^{-5}	0.995877
12346	1	0.000172	0.989549	1.370789	0.241685
12356	1	0.192403	0.660927	2.708191	0.099844
12456	1	-7.6×10^{-5}	1.000000	1.53×10^{-5}	0.996883
13456	1	-7.2×10^{-5}	1.000000	0.000538	0.981497
13456	1	-7.2×10^{-5}	1.000000	0.000538	0.981497

Zależności cząstkowe (jeżeli są istotne) wskazują, że określona interakcja ma wpływ na lepsze dopasowanie modelu w obecności wszystkich innych rozważanych interakcji. Jest ona szacowana na podstawie porównania dwóch hipotetycznych modeli: w jednym z nich występują wszystkie interakcje rozpatrywanego rzędu włącznie z tą badaną, w drugim modelu występują natomiast wszystkie interakcje danego rzędu oprócz tej, którą analizujemy. Jeżeli różnica między takimi dwoma modelami jest istotna, to znaczy, że badana interakcja przyczynia się istotnie do poprawienia dopasowania modelu i dlatego zostawiamy tą interakcję w naszym modelu. Na podstawie istotności zależności cząstkowych weryfikujemy wkład (kontrybucję) każdej z rozpatrywanych zależności w dopasowanie modelu, który ma jak najlepiej opisywać badane zjawisko (w naszym przykładzie: wystąpienie zawału). Liczenie zależności cząstkowych jest procedurą analogiczną do analizy krokowej wstecznej z kolejnym usuwaniem zmiennych.

Zależności brzegowe służą do porównania wpływu interakcji w procedurze analogicznej do analizy krokowej postępującej: porównujemy model bez żadnych interakcji rozpatrywanego rzędu z modelem zawierającym jako jedyną tę interakcję, którą badamy. Wynik analizy mówi nam, czy dana interakcja ma jakikolwiek wpływ na polepszenie dopasowania modelu w sytuacji, gdy nie ma w modelu jeszcze żadnych innych interakcji. Jeżeli współzależność brzegowa pozostaje istotna, zaś cząstkowa nieistotna, mamy do czynienia z sytuacją analogiczną jak w metodzie regresji wielokrotnej: istotną korelację liniową Pearsona jakiejś zmiennej i nieistotną korelację cząstkową tej zmiennej. Odzwierciedla to sytuację, gdy wpływ danej

zmiennej jest w pełni tłumaczony przez inne zmienne. Analogicznie, w naszym modelu – gdybyśmy uwzględnili pozostałe interakcje tego rzędu co badana, wtedy ta rozważana interakcja przestaje być istotna i nie musimy jej uwzględniać w modelu.

Oglądając istotności w tabeli powyżej widzimy, że powinniśmy włączyć do modelu następujące interakcje:
– zależność między zawałem (1) a płcią (3) (interakcja 13),
– zależność między zawałem (1) a polimorfizmem VNTR (4) i polimorfizmem HPA2 (5) (interakcja 145).

Zgodnie z interpretacją podaną powyżej, zależność między zawałem (1) a polimorfizmem VNTR (4) i polimorfizmem ^{807}C/T (6) (interakcja 146) jest całkowicie wyjaśniana przez inne interakcje.

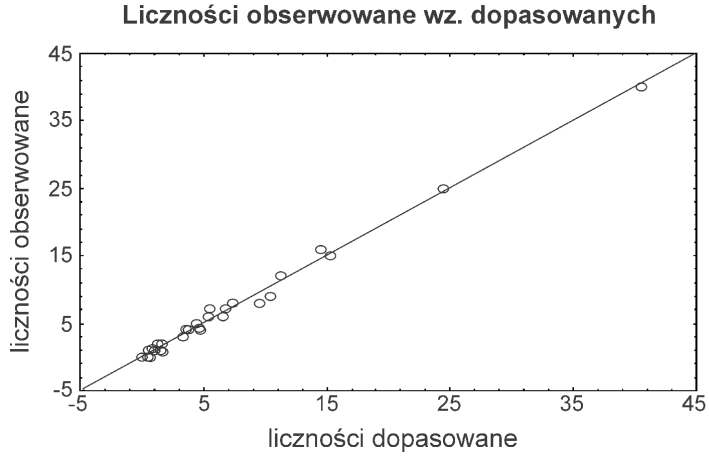
Z kolei, istotne zależności między kategorią wieku (2) a płcią (3) (interakcja 23), jak również między polimorfizmem VNTR (4) i polimorfizmem HPA2 (5) nie są dla nas ciekawe lub są oczywiste i nie ma potrzeby ich badać.

Analiza tak dobranego modelu daje wartości testów chi^2 nieistotne statystycznie (testy te oceniają dobroć dopasowania modelu do analizowanych danych; jeżeli są istotne to model odrzucamy):

chi^2 największej wiarygodności	***d.f.***	**istotność**
6.891	23	0.9995

chi^2 Pearsona	***d.f.***	**istotność**
5.805	23	0.9999

Na podstawie tych wartości możemy uznać, że model wybrany przez nas opisuje w sposób zadowalający obserwacje w próbie. Dodatkowym potwierdzeniem tego wniosku jest wykres zależności wartości obserwowanych i wartości dopasowanych, na którym brak jest obserwacji odstających.



Zestawiamy tabele brzegowe liczebności.

Iloraz szans i przedział ufności (*CI*): zawał *vs.* płeć.

	zawał			
płeć	**0**	**1**		
0	38	20		
1	57	100		
chi^2	istotność *p*	OR	–95% *CI*	+95% *CI*
14.615	0.001	3.333	1.798	6.180

Iloraz szans i przedział ufności (*CI*): zawał *vs.* wiek.

	zawał			
wiek	**0**	**1**		
0	64	77		
1	31	43		
chi^2	istotność *p*	OR	–95% *CI*	+95% *CI*
0.267	*n.s.*	1.153	0.672	1.978

Iloraz szans i przedział ufności (*CI*): zawał *vs.* VNTR-B.

	zawał			
VNTR-B(+)	**0**	**1**		
0	75	96		
1	20	24		
chi^2	istotność *p*	OR	–95% *CI*	+95% *CI*
0.061	*n.s.*	0.938	1.567	0.561

Iloraz szans i przedział ufności (*CI*): zawał *vs.* HPA-2b.

	zawał			
HPA-2b	**0**	**1**		
0	75	97		
1	20	23		
chi^2	istotność *p*	OR	–95% *CI*	+95% *CI*
0.139	*n.s.*	0.889	1.648	0.480

Iloraz szans i przedział ufności (*CI*): zawał *vs.* ^{807}T(+).

	zawał			
^{807}T(+)	**0**	**1**		
0	27	41		
1	68	79	67	
chi^2	istotność *p*	OR	–95% *CI*	+95% *CI*
0.843	*n.s.*	0.765	1.355	0.432

Iloraz szans i przedział ufności (*CI*): zawał *vs.* VNTR-B(+) w grupie: ^{807}CC.

	zawał			
VNTR-B(+)	**0**	**1**		
0	23	29		
1	4	12		
chi^2	istotność *p*	OR	–95% *CI*	+95% *CI*
2.063	*n.s.*	2.379	0.729	7.765

Iloraz szans i przedział ufności (*CI*): zawał *vs.* VNTR-B(+) w grupie: ^{807}T(+).

	zawał			
VNTR-B(+)	**0**	**1**		
1	16	12		
0	52	67		
chi^2	istotność *p*	OR	–95% *CI*	+95% *CI*
1.668	*n.s.*	1.718	0.756	3.906

W podsumowaniu, stwierdzamy, że analiza wartości (liczebności) brzegowych wskazuje, iż:
- wśród mężczyzn zawały obserwuje się istotnie częściej (64% *vs.* 34%); iloraz szans wynosi 3.33 (*95%CI* 1.80-6.18. $p<0.001$), co oznacza że u mężczyzn ryzyko wystąpienia zawału jest ponad 3-krotnie wyższe niż u kobiet;
- żaden z badanych genotypów, tzn. ani VNTR-B (20% w grupie z zawałem *vs.* 21% w grupie bez zawału), ani HPA2b (19% w grupie z zawałem *vs.* 21% w grupie bez zawału), ani ^{807}T(+) (66% w grupie z zawałem *vs.* 72% w grupie bez zawału) nie jest związany z istotnie wyższym ryzykiem zawału.

Przykład 93

Wśród pacjentów z chorobą niedokrwienną serca przeprowadzono wybrane badania biochemiczne (lipidy surowicy krwi), badania genetyczne (polimorfizmy wybranych glikoprotein płytek krwi) oraz zebrano wywiad o każdym pacjencie.

Celem badania było znalezienie czynników, które najsilniej wpływają na wystąpienie zawału, a przede wszystkim określenie, czy występowanie polimorfizmu glikoproteiny płytkowej IIIa: Pl^{A2}(+) (HPA-1b) ma wpływ na częstość występowania zawału mięśnia sercowego.

Badano wpływ:

polimorfizmu HPA-2b: 0 – HPA-2b (–), 1 – HPA-2b (+)
polimorfizmu HPA-1: 0 – Pl^{A2}(–), 1 – Pl^{A2}(+),
polimorfizmu ^{807}C/T: 0 – ^{807}T (–), 1 – ^{807}T (+),
palenia: 0 – nie pali, 1 – pali,
dyslipidemii: 0 – brak, 1 – jest,

na wystąpienie zawału (1). Występowanie zawału było więc zmienną zależną od występowania pozostałych parametrów (zmienne niezależne). Badaniami objęto 209 pacjentów (wyniki w arkuszu ***analiza log-liniowa.xls***).

W celu wybrania modelu i wyszczególnienia potencjalnych interakcji między zmiennymi przeprowadzono jednoczesne testy, że wszystkie interakcje między k czynnikami są jednocześnie równe 0. Wyniki tych testów przedstawia tabela:

Wyniki dotyczące wszystkich potencjalnych interakcji k czynników. Jednoczesne testy, że wszystkie interakcje k czynników są jednocześnie równe 0.

interakcje stopnia	stopnie swobody	chi^2 najw. wiarygodn.	istotność p	chi^2 Pearsona	istotność p
1	6	177.49	**0.000**	223.11	**0.000**
2	15	25.78	**0.040**	31.78	**0.007**
3	20	17.40	0.627	14.66	0.795
4	15	13.47	0.566	14.09	0.519
5	6	5.55	0.475	5.98	0.426
6	1	0.49	0.482	0.50	0.480

Modele bez interakcji dwuczynnikowych (p = 0.007) należy odrzucić. Dołączenie interakcji rzędu 3 i wyższego nie daje istotnej poprawy dopasowania modelu (p > 0.6). Wskazuje to, że model najmniej złożony to model bez interakcji trójwymiarowych i wyższych, natomiast występują interakcje rzędu drugiego, dlatego też dołączamy je do modelu. Które z nich są istotne, wskazują wyniki w tabeli zależności brzegowych i zależności cząstkowych:

Testy zależności brzegowych i cząstkowych.

	stopnie swobody	zal. cząstkowe chi^2	zal. cząstkowe p	zal. brzegowe chi^2	zal. brzegowe p
1	1	32.12706	1.46×10^{-8}	32.12706	1.46×10^{-8}
2	1	16.66175	4.48×10^{-5}	16.66175	4.48×10^{-5}
3	1	25.03827	5.65×10^{-7}	25.03827	5.65×10^{-7}
4	1	2.59808	0.107003	2.59808	0.107003
5	1	65.8327	5.08×10^{-16}	65.8327	5.08×10^{-16}
6	1	35.22789	2.96×10^{-9}	35.22789	2.96×10^{-9}
12	1	0.05658	0.811988	0.199738	0.654936
13	1	0.064682	0.799245	0.054398	0.815582
14	1	0.230618	0.63107	0.1754	0.67536
15	1	1.599953	0.205919	1.73819	0.187378
16	1	0.002708	0.958495	3.05×10^{-5}	0.995592
23	1	2.345448	0.125659	2.246918	0.133891
24	1	1.343548	0.246418	1.726379	0.188883
25	1	5.65332	0.017428	4.904739	0.02679
26	1	0.7985	0.371549	0.240646	0.623743
34	1	6.451653	0.01109	7.853821	0.005074
35	1	1.33799	0.247397	0.843613	0.358373
36	1	3.691631	0.054695	3.772797	0.052101
45	1	0.321602	0.570651	0.220566	0.638612
46	1	0.589993	0.442428	1.024933	0.311359
56	1	1.242462	0.265005	0.717285	0.397042
123	1	0.235624	0.627388	0.022644	0.880387
124	1	0.478254	0.489219	0.975769	0.323253
125	1	0.916971	0.338279	0.831001	0.36199
126	1	0.151594	0.697019	0.016098	0.899038
134	1	0.236856	0.626489	0.273926	0.600714
135	1	0.562737	0.453165	0.325897	0.568089
136	1	0.013065	0.908998	0.215622	0.642399
145	1	0.775198	0.37862	0.928894	0.335158

146	1	0.17955	0.671763	0.269348	0.603773
156	1	4.300642	0.038106	3.61805	0.057165
234	1	0.334557	0.562992	0.001984	0.964476
235	1	4.908098	0.026738	4.704231	0.030096
236	1	0.336409	0.561914	0.016266	0.898516
245	1	0.454571	0.500177	1.011475	0.314557
246	1	0.604738	0.436782	0.313965	0.575261
256	1	1.368301	0.242113	0.702591	0.401921
345	1	0.29142	0.589316	0.076187	0.782534
346	1	2.189228	0.138989	1.88501	0.169775
356	1	0.002539	0.959816	0.038155	0.845134
456	1	0.151415	0.69719	0.395523	0.529415
1234	1	0.657006	0.417625	0.089981	0.764203
1235	1	0.094532	0.758494	0.015923	0.899587
1236	1	2.205177	0.137558	1.596031	0.206476
1245	1	0.110759	0.739284	0.042213	0.837215
1246	1	–0.01391	1.000000	0.501877	0.478682
1256	1	2.036347	0.153588	1.624916	0.202417
1345	1	2.319882	0.12774	2.225838	0.135729
1346	1	0.583974	0.444765	1.304855	0.253337
1356	1	2.836885	0.092132	1.320717	0.250472
1456	1	2.476865	0.115542	2.109856	0.146363
2345	1	–0.04	1.000000	0.169205	0.680822
2346	1	1.37298	0.241309	0.508865	0.475636
2356	1	0.963897	0.326214	0.457214	0.498934
2456	1	–0.07736	1.000000	9.92×10^{-5}	0.992054
3456	1	0.594497	0.440691	0.585098	0.444326
12345	1	1.087157	0.29711	1.098335	0.29464
12346	1	0.102675	0.748645	0.504623	0.477481
12356	1	0.503274	0.47807	0.586803	0.443664
12456	1	3.295208	0.069492	3.501484	0.061323
13456	1	0.077801	0.780302	0.019821	0.888039
13456	1	0.010275	0.919259	0.378567	0.538376

Oglądając istotności w tabeli powyżej widzimy, że powinniśmy włączyć do modelu następujące interakcje:
- zależność między zawałem (4) a paleniem (3) (interakcja 34),
- zależność między dyslipidemią (2) a polimorfizmem HPA-2 (5) (interakcja 25),
- zależność między polimorfizmem HPA-1 (1) a polimorfizmem HPA-2 (5) i polimorfizmem ^{807}C/T (6) (interakcja 156),
- zależność między polimorfizmem HPA-2 (5), dyslipoproteinemią (2) i paleniem (3) (interakcja 235).

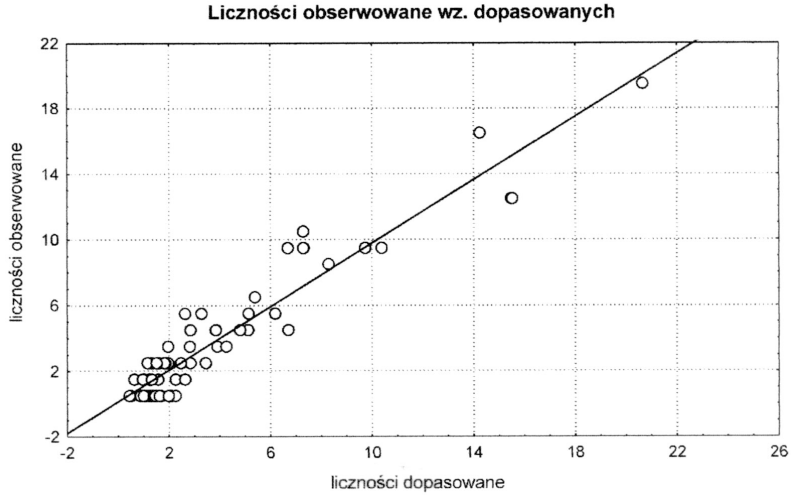
Ostatecznie testujemy model obejmujący powyższe zależności, a ponieważ szczególnie interesuje nas wpływ zmiennej HPA-1 (1) na występowanie zawału (4), dodatkowo testujemy także zależności:
- zależność między zawałem (4) a polimorfizmem HPA-1 (1) (interakcja 14),
- zależność między polimorfizmem HPA-1 (1), paleniem (3) a zawałem (4) (interakcja 134),
- zależność między polimorfizmem HPA-1 (1), dyslipoproteinemią (2) i zawałem (4) (interakcja 124).

Analiza tak dobranego modelu daje wartości testów chi^2 nieistotne statystycznie. Pamiętając, że testy te oceniają dobroć dopasowania modelu do analizowanych danych oraz że istotne wartości statystyki chi^2 upoważniają nas do odrzucenia modelu, przeprowadzamy test chi^2 największej wiarygodności oraz test chi^2 Pearsona:

chi^2 największej wiarygodności	***d.f.***	**istotność**
33.043	42	0.837

chi^2 Pearsona	***d.f.***	**istotność**
31.84	42	0.872

Na podstawie tych wartości możemy uznać, że model wybrany przez nas opisuje w sposób zadowalający obserwacje w próbie. Dodatkowym potwierdzeniem jest wykres zależności wartości obserwowanych i wartości dopasowanych, na którym brak jest obserwacji odstających.



Zestawiamy tabele brzegowe liczebności.

Iloraz szans i przedział ufności (*CI*): palenie *vs.* zawał.

	palenie			
zawał	**0**	**1**		
0	47	61		
1	35	98		
chi^2	istotność *p*	OR	–95% *CI*	+95% *CI*
7.851	0.01	2.157	1.260	3.694

Iloraz szans i przedział ufności (*CI*): dyslipidemia *vs.* HPA–2b.

	dyslipidemia			
HPA-2b	**0**	**1**		
0	122	60		
1	30	29		
chi^2	istotność *p*	OR	–95% *CI*	+95% *CI*
4.943	0.05	1.966	1.083	3.566

Iloraz szans i przedział ufności (*CI*): HPA-1b *vs*. HPA-2b w grupie: ^{807}C/T (0).

	HPA-1b			
HPA-2b	**0**	**1**		
0	41	13		
1	10	11		
chi^2	istotność *p*	OR	–95% *CI*	+95% *CI*
5.438	0.05	3.469	1.219	9.870

Iloraz szans i przedział ufności (*CI*): HPA-1b *vs*. HPA-2b w grupie: ^{807}C/T (1).

	HPA-1b			
HPA-2b	**0**	**1**		
0	87	41		
1	26	12		
chi^2	Istotność *p*	OR	–95% *CI*	+95% *CI*
0.046	*n.s.*	0.979	1.185	0.810

Iloraz szans i przedział ufności (*CI*): HPA-1b *vs*. zawał.

	HPA-1b			
zawał	**0**	**1**		
0	75	33		
1	89	44		
chi^2	istotność *p*	OR	–95% *CI*	+95% *CI*
0.199	*n.s.*	1.124	0.673	1.876

Iloraz szans i przedział ufności (*CI*): HPA-1b *vs*. zawał w grupie: palenie (0).

	HPA-1b			
palenie	**0**	**1**		
0	33	14		
1	22	13		
chi^2	istotność *p*	OR	–95% *CI*	+95% *CI*
0.531	*n.s.*	1.393	0.571	3.395

Iloraz szans i przedział ufności (*CI*). HPA-1b *vs*. zawał w grupie: palenie (1).

	HPA-1b			
palenie	**0**	**1**		
0	42	19		
1	67	31		
chi^2	istotność *p*	OR	–95% *CI*	+95% *CI*
0.037	*n.s.*	1.023	0.813	1.287

Iloraz szans i przedział ufności (*CI*): HPA-1b *vs*. zawał w grupie: dyslipidemia (0).

	HPA-1b			
dyslipidemia	**0**	**1**		
0	53	20		
1	52	27		
chi^2	istotność *p*	OR	*–95% CI*	*+95% CI*
0.852	*n.s.*	1.376	0.699	2.710

Iloraz szans i przedział ufności (*CI*): HPA-1b *vs*. zawał w grupie: dyslipidemia (1).

	HPA-1b			
dyslipidemia	**0**	**1**		
0	22	13		
1	37	17		
chi^2	istotność *p*	OR	*–95% CI*	*+95% CI*
0.340	*n.s.*	0.778	1.812	0.334

Analiza wartości (liczebności) brzegowych wskazuje, że:

– wśród osób palących zawały obserwuje się istotnie częściej (62% *vs*. 43%); iloraz szans wynosi 2.16 (*CI*95% 1.26-3.70 $p<0.01$), co oznacza że u „palaczy" ryzyko wystąpienia zawału jest ponad 2-krotnie wyższe niż u osób niepalących;

– występowanie polimorfizmu HPA-1b glikoproteiny IIIa nie jest związane z istotnie wyższym ryzykiem zawału (54% w grupie HPA-1a *vs*. 57% w grupie HPA-1b); OR = 1.124 (*CI*95% 0.67-1.88, *n.s.*), co oznacza, że występowanie tego polimorfizmu nie może być jednoznacznie postrzegane jako czynnik ryzyka zawału. Dane wskazują, że o istotności polimorfizmu HPA-1 jako czynnika zawału nie decydują także inne czynniki, takie jak palenie tytoniu czy dyslipidemia (OR = 1.39 u niepalących i OR = 1.02 u palących, oraz OR = 1.38 u pacjentów bez oznak dyslipidemii i OR = 0.78 u pacjentów z występującą dyslipidemią).

– dyslipidemia występuje istotnie częściej w grupie pacjentów z ChNS, którzy są nosicielami polimorfizmu HPA-2b (39% *vs*. 33%, OR = 1.97 (*CI*95% 1.08-3.57, $p<0.05$)

– polimorfizm HPA-1b występuje istotnie częściej u nosicieli polimorfizmu HPA-2b niż HPA-2a (52% *vs*. 24%, OR = 3.47, $p < 0.05$), ale tylko wtedy, gdy osoby te są jednocześnie nosicielami polimorfizmu ^{807}CC; zależność taka nie występuje natomiast u nosicieli ^{807}T(+).

Rozdział 24

Zastosowania metod nieparametrycznych

Testy do porównania dwóch prób

Test Manna-Whitneya

Przykład 94

Czy stężenie PAI-1 w osoczu krwi (ng/ml) w grupie pacjentów z cukrzycą typu 2 jest istotnie różne od stężenia PAI-1 w osoczu zdrowych ochotników?

kontrola		cukrzyca	
50.98	12.42	16.32	57.16
8.49	19.04	41.28	5.81
10.14	69.69	70.55	27.41
26.28	11.52	7.99	55.98
43.76	45.54	26.58	21.32
36.28	14.93	16.41	58.62
46.16	23.52	59.22	37.61
8.10	16.66	18.54	50.14
19.10	54.31	45.15	64.20
11.85	12.35	31.99	39.49
21.53	35.93		
38.28	25.70		
48.07	9.63		
14.68	52.23		
27.31	36.42		
82.89			

Chcąc policzyć istotność różnic między dwiema zmiennymi testem Manna-Whitneya, pierwszym krokiem powinno być posortowanie wyników i nadanie im rang. Przyjmijmy, że grupa 1 będzie oznaczała osoby „kontrolne", zaś grupa 2 – osoby chore na cukrzycę. Tabelę zestawiono w taki sposób, że trzy kolumny po lewej stronie zawierają wyniki uporządkowane rosnąco bez względu na przynależność do grupy, natomiast trzy prawe kolumny zawierają wyniki uporządkowane rosnąco osobno dla każdej grupy.*

ranga	PAI-1	grupa	ranga	PAI-1	grupa
1	**5.81**	**2**	3	8.10	1
2	**7.99**	**2**	4	8.49	1
3	8.10	1	5	9.63	1
4	8.49	1	6	10.14	1

5	9.63	1	7	11.52	1
6	10.14	1	8	11.85	1
7	11.52	1	9	12.35	1
8	11.85	1	10	12.42	1
9	12.35	1	11	14.68	1
10	12.42	1	12	14.93	1
11	14.68	1	15	16.66	1
12	14.93	1	17	19.04	1
13	**16.32**	**2**	18	19.10	1
14	**16.41**	**2**	20	21.53	1
15	16.66	1	21	23.52	1
16	**18.54**	**2**	22	25.70	1
17	19.04	1	23	26.28	1
18	19.10	1	25	27.31	1
19	**21.32**	**2**	28	35.93	1
20	21.53	1	29	36.28	1
21	23.52	1	30	36.42	1
22	25.70	1	32	38.28	1
23	26.28	1	35	43.76	1
24	**26.58**	**2**	37	45.54	1
25	27.31	1	38	46.16	1
26	**27.41**	**2**	39	48.07	1
27	**31.99**	**2**	41	50.98	1
28	35.93	1	42	52.23	1
29	36.28	1	43	54.31	1
30	36.42	1	49	69.69	1
31	**37.61**	**2**	51	82.89	1
32	38.28	1	**1**	**5.81**	**2**
33	**39.49**	**2**	**2**	**7.99**	**2**
34	**41.28**	**2**	**13**	**16.32**	**2**
35	43.76	1	**14**	**16.41**	**2**
36	**45.15**	**2**	**16**	**18.54**	**2**
37	45.54	1	**19**	**21.32**	**2**
38	46.16	1	**24**	**26.58**	**2**
39	48.07	1	**26**	**27.41**	**2**
40	**50.14**	**2**	**27**	**31.99**	**2**
41	50.98	1	**31**	**37.61**	**2**
42	52.23	1	**33**	**39.49**	**2**
43	54.31	1	**34**	**41.28**	**2**
44	**55.98**	**2**	**36**	**45.15**	**2**
45	**57.16**	**2**	**40**	**50.14**	**2**
46	**58.62**	**2**	**44**	**55.98**	**2**
47	**59.22**	**2**	**45**	**57.16**	**2**
48	**64.20**	**2**	**46**	**58.62**	**2**
49	69.69	1	**47**	**59.22**	**2**
50	**70.55**	**2**	**48**	**64.20**	**2**
51	82.89	1	**50**	**70.55**	**2**

* wartości dla grupy pacjentów z cukrzycą pogrubiono.

Zestawmy wyniki dla obu grup obok siebie.

ranga	**PAI-1**	**kontrola**	**ranga**	**PAI-1**	**cukrzyca**
3	8.10	1	1	5.81	2
4	8.49	1	2	7.99	2
5	9.63	1	13	16.32	2
6	10.14	1	14	16.41	2
7	11.52	1	16	18.54	2
8	11.85	1	19	21.32	2
9	12.35	1	24	26.58	2

10	12.42	1	26	27.41	2
11	14.68	1	27	31.99	2
12	14.93	1	31	37.61	2
15	16.66	1	33	39.49	2
17	19.04	1	34	41.28	2
18	19.10	1	36	45.15	2
20	21.53	1	40	50.14	2
21	23.52	1	44	55.98	2
22	25.70	1	45	57.16	2
23	26.28	1	46	58.62	2
25	27.31	1	47	59.22	2
28	35.93	1	48	64.20	2
29	36.28	1	50	70.55	2
30	36.42	1			
32	38.28	1	n =	20.00	
35	43.76	1	**ΣR_2 =**	**596.00**	
37	45.54	1			
38	46.16	1			
39	48.07	1			
41	50.98	1			
42	52.23	1			
43	54.31	1			
49	69.69	1			
51	82.89	1			
n_1 =	31.00				
ΣR_1 =	**730.00**				

Wartość statystyki testu Manna-Whitneya wynosi:

$$U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - \sum R_{1j} = (31)(20) + \frac{31(32)}{2} - 730 = 1116 - 730 = 386$$

$$U' = n_2 n_1 + \frac{n_2(n_2+1)}{2} - \sum R_{2j} = (31)(20) + \frac{20(21)}{2} - 596 = 234$$

dla sprawdzenia:

$$U = n_1 n_2 - U' = 620 - 234 = 386$$

$$z = \frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 [n_1 + (n_2+1)]}{12}}}$$

$$\frac{386 - \frac{(31)(20)}{2}}{\sqrt{\frac{(31)(20)[31+(20+1)]}{12}}} = \frac{386-310}{51.833} = \frac{76}{51.833} = -1.46$$

Nie mamy podstaw do odrzucenia hipotezy zerowej, gdyż $|z| < z_{\alpha/2} = 1.96$. Wnioskujemy zatem, że stężenia PAI-1 w osoczu krwi pełnej u zdrowych ochotników i pacjentów z cukrzycą typu 2 nie są istotnie różne.

Test sumy rang Wilcoxona

Przykład 95

W grupie 29 kobiet ciężarnych 15 matek miało cukrzycę typu ciężarnych. Czy masa urodzeniowa noworodków jest różna w grupie matek z cukrzycą i w grupie matek zdrowych?

masa	cukrzyca	masa	kontrola
3.51	2	3.59	1
3.73	2	3.52	1
3.99	2	2.9	1
3.83	2	2.76	1
3.31	2	3.18	1
4.08	2	3.27	1
2.71	2	2.38	1
3.54	2	3.6	1
3.79	2	2.34	1
3.6	2	2.84	1
3.26	2	3.85	1
4.13	2	3.23	1
3.21	2	3.63	1
3.61	2	3.75	1
3.6	2		

Posłużymy się testem sumy rang Wilcoxona. Test ten przeprowadza się następująco:

1. Obserwacje dla obu grup razem porządkujemy w kolejności rosnącej i nadajemy im rangi; dla identycznych wartości obliczamy rangi wiązane.
2. Sumujemy rangi w grupie o mniejszej liczebności: T = suma rang dla mniej licznej grupy; jeżeli obie grupy mają tą samą liczebność wybieramy dowolną z nich.
3. Obliczoną wartość T porównujemy z wartością tablic dla testu sumy rang Wilcoxona: aby odrzucić hipotezę zerową, nasza obliczona wartość T musi znajdować się poza zakresem nieistotnych sum w tablicach.

Sortujemy dane i nadajemy im rangi:

ranga	masa	grupa	ranga	masa	grupa
1	2.34	1	16	3.59	1
2	2.38	1	17	***3.60***	1
3	2.71	2	18	***3.60***	2
4	2.76	1	19	***3.60***	2
5	2.84	1	20	3.61	2
6	2.90	1	21	3.63	1
7	3.18	1	22	3.73	2
8	3.21	2	23	3.75	1
9	3.23	1	24	3.79	2
10	3.26	2	25	3.83	2

11	3.27	1	26	3.85	1
12	3.31	2	27	3.99	2
13	3.51	2	28	4.08	2
14	3.52	1	29	4.13	2
15	3.54	2			

Zauważmy, że wystąpiły trzy powtórzenia wartości 3.60; w porządku rosnącym przyporządkowano im rangi 17, 18 i 19, z których obliczamy rangę wiązaną jako średnią (17+18+19)/3 = 18

Nasza uporządkowana tabela z danymi ma postać:

ranga	**masa**	**grupa**	**ranga**	**masa**	**grupa**
1	2.34	1	3	2.71	2
2	2.38	1	8	3.21	2
4	2.76	1	10	3.26	2
5	2.84	1	12	3.31	2
6	2.90	1	13	3.51	2
7	3.18	1	15	3.54	2
9	3.23	1	**18**	***3.60***	2
11	3.27	1	**18**	***3.60***	2
14	3.52	1	20	3.61	2
16	3.59	1	22	3.73	2
18	***3.60***	1	24	3.79	2
21	3.63	1	25	3.83	2
23	3.75	1	27	3.99	2
26	3.85	1	28	4.08	2
			29	4.13	2

Dla grupy kobiet bez cukrzycy mamy:

$$n_1 = 14 \quad \textit{suma rang}\ (R_1) = 163$$

W grupie kobiet z cukrzycą ciężarnych:

$$n_2 = 15 \quad \textit{suma rang}\ (R_2) = 272$$

Nasza suma rang dla mniejszej grupy to $R_1 = 163$.

Sprawdzamy wartości krytyczne dla testu sumy rang Wilcoxona: dla liczebności grup 14 i 15 i dla $\alpha = 0.05$ zakres 164-256 odpowiada wartościom nieistotnym, tzn. wartości poniżej 164 oraz powyżej 256 są istotne z prawdopodobieństwem 95%. Zauważmy, że gdyby grupy były jednakowo liczne i wybralibyśmy $R_2 = 272$ jako wartość statystyki testu, to wynik byłby także istotny statystycznie.

Dla tych samych danych policzmy jeszcze wartość statystyki *z* testu normalnego:

$$U = R_2 - \frac{n_2(n_2+1)}{2} = 272 - \frac{15(15+1)}{2} = 152$$

oraz

$$U' = R_1 - \frac{n_1(n_1+1)}{2} = 163 - \frac{14(14+1)}{2} = 58$$

$$z = \frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 [n_1 + (n_2 + 1)]}{12}}} = \frac{152 - \frac{(14)(15)}{2}}{\sqrt{\frac{(14)(15)[14 + (15 + 1)]}{12}}} = \frac{47}{22.913} = 2.051$$

Dla z = 2.05 wartość prawdopodobieństwa wynosi $\frac{\alpha}{2} = 0.02018$, czyli z prawdopodobieństwem prawie 96% (dla testu obustronnego – gdyż testujemy, że masa jest różna, czyli albo większa albo mniejsza – α = 2 x 0.02018 = 0.04036, 1 – 0.0404 = 0.9596) możemy odrzucić hipotezę zerową mówiąca, że masa urodzeniowa noworodków w grupie kobiet z cukrzycą ciężarnych jest taka sama jak masa urodzeniowa noworodków w grupie kobiet zdrowych

Test mediany dla dwóch prób

Przykład 96

Zastosuj test mediany dla dwóch prób do danych z poprzedniego przykładu.

- H_0: mediany (*Me*) są takie same w obu próbach, czyli około połowa wszystkich przypadków w każdej z prób wypada powyżej, a połowa poniżej wspólnej mediany
- H_A: mediany nie są takie same w obu próbach

kontrola	cukrzyca	powyżej Me	nie powyżej Me
8.10	5.81		
8.49	7.99		
9.63	16.32		
10.14	16.41		
11.52	18.54		
11.85	21.32		
12.35	26.58		
12.42	27.41		
14.68	31.99		TAK
14.93	37.61		TAK
16.66	39.49		TAK
19.04	41.28		TAK
19.10	45.15		TAK
21.53	50.14		TAK
23.52	55.98		TAK
25.7	57.16		TAK
26.28	58.62		TAK
27.31	59.22		TAK
35.93	64.2	TAK	TAK
36.28	70.55	TAK	TAK
36.42		TAK	
38.28		TAK	
43.76		TAK	
45.54		TAK	
46.16		TAK	
48.07		TAK	
50.98		TAK	
52.23		TAK	

	54.31		TAK
	69.69		TAK
	82.89		TAK
Me:	25.70	38.55	
n:	*31*	*20*	
wspólna Me: **27.41**			

Sprawdzamy rozkład wartości powyżej i nie powyżej wartości wspólnej mediany:

	kontrola	**cukrzyca**	***razem***
powyżej mediany	13	12	**25**
nie powyżej mediany	18	8	**26**
razem	**31**	**20**	**51**

Obliczamy prawdopodobieństwo, że obserwowane różnice w rozkładach wartości mediany mogłyby wystąpić przez czysty przypadek:

$$p = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{n!\,a!\,b!\,c!\,d!} = \frac{(25!)(26!)(31!)(20!)}{(51!)(13!)(12!)(18!)(8!)} = 0.105$$

Hipotezy zerowej nie odrzucamy, ponieważ prawdopodobieństwo, że obserwowane różnice w rozkładach wartości mediany mogłyby wystąpić przez czysty przypadek wynosi ponad 10%.

Obliczmy jeszcze dla tego samego przykładu wartość statystyki chi^2:

$$\chi^2 = \frac{n\left(|ad-bc|-\frac{n}{2}\right)^2}{(a+c)(b+d)((a+b)(c+d)} = \frac{51\left(|(13*8)-(12*18)|-\frac{51}{2}\right)^2}{(31)(20)(25)(26)} = \frac{3815948}{403000} = 0.947$$

$$\text{dla} \qquad df = (k-1)(2-1) = k-1 = 2-1 = 1, \quad \chi^2_{0.05,1} = 3.841,$$

czyli nie mamy podstaw, aby odrzucić hipotezę zerową mówiącą, że mediany są takie same w obu próbach, czyli około połowa wszystkich przypadków w każdej z prób wypada powyżej, a połowa poniżej wspólnej mediany.

Test Kolmogorova-Smirnova

Został omówiony poniżej, w części *„Badanie dopasowania rozkładu"*.

Testy do porównania więcej niż dwóch prób

Test Kruskala-Wallisa

Przykład 97

Badano wpływ 5 różnych antybiotyków na zahamowanie wzrostu kultur bakteryjnych. Skuteczność działania antybiotyków oceniano na podstawie liczby „łysinek" powstających w hodowlach na szalkach Petriego.

antybiotyk 1	antybiotyk 2	antybiotyk 3	antybiotyk 4	antybiotyk 5
13	5	10	5	12
7	4	11	6	9
4	3	12	2	12
8	10	14	6	6
11	8	12	8	9
6	5	15	5	7
4	2	8	5	12
8	6	11	1	15
13	9	7	7	14
6	6	12	5	11
21	9	7	4	4
14	7	13	3	16
6	6	11	5	7
10	12	12	6	10
11	9	12	7	8
7	11	13	3	7
5	10	6	3	11
2	9	9	5	11
6	7	14	6	5
7	7	8	7	5

Czy testowane antybiotyki różnią się skutecznością niszczenia kultur bakterii, tzn. czy liczba „łysinek" powstających w obecności każdego z badanych antybiotyków była istotnie różna?

Z uwagi na porządkowy charakter zmiennych wykorzystamy analizę wariancji Kruskala-Wallisa.

Ustalamy rangi i ich rozdział do poszczególnych grup oraz obliczamy sumy rang w każdej grupie:

antybiotyk 1		antybiotyk 2		antybiotyk 3		antybiotyk 4		antybiotyk 5	
2	**3**	2	**3**	6	**31**	1	1	4	**11**
4	**11**	3	**6.5**	7	**44.5**	2	**3**	5	**19**
4	**11**	4	**11**	7	**44.5**	3	**6.5**	5	**19**
5	**19**	5	**19**	8	**55**	3	**6.5**	6	**31**
6	**31**	5	**19**	8	**55**	3	**6.5**	7	**44.5**
6	**31**	6	**31**	9	**62**	4	**11**	7	**44.5**
6	**31**	6	**31**	10	**68**	5	**19**	7	**44.5**
6	**31**	6	**31**	11	**75**	5	**19**	8	**55**
7	**44.5**	7	**44.5**	11	**75**	5	**19**	9	**62**
7	**44.5**	7	**44.5**	11	**75**	5	**19**	9	**62**
7	**44.5**	7	**44.5**	12	**84**	5	**19**	10	**68**

8	**55**	8	**55**	12	**84**	5	**19**	11	**75**
8	**55**	9	**62**	12	**84**	6	**31**	11	**75**
10	**68**	9	**62**	12	**84**	6	**31**	11	**75**
11	**75**	9	**62**	12	**84**	6	**31**	12	**84**
11	**75**	9	**62**	13	**90.5**	6	**31**	12	**84**
13	**90.5**	10	**68**	13	**90.5**	7	**44.5**	12	**84**
13	**90.5**	10	**68**	14	**94.5**	7	**44.5**	14	**94.5**
14	**94.5**	11	**75**	14	**94.5**	7	**44.5**	15	**97.5**
21	100	12	**84**	15	**97.5**	8	**55**	16	99
ΣR_1	***1005***	**ΣR_2**	***883***	**ΣR_3**	***1472.5***	**ΣR_4**	***461***	**ΣR_5**	***1228.5***

Na podstawie rang dla k grup (poziomów) obliczamy wartość statystyki H testu Kruskala-Wallisa:

$$H = \frac{12}{N(N+1)}\left[\sum \frac{\left(\sum R_{ij}\right)^2}{n_j}\right] - 3(N+1)$$

$$H = \frac{12}{100(100+1)}\left[\frac{(1005)^2}{20} + \frac{(883)^2}{20} + \frac{(1472.5)^2}{20} + \frac{(461)^2}{20} + \frac{(1228.5)^2}{20}\right] - 3(100+1) = 34.41$$

Obliczamy poprawkę dla rang wiązanych:

$$C = 1 - \left[\frac{\Sigma\left(t^3 - t\right)}{N^3 - N}\right] =$$

$$= \left[\frac{1[(2)^3 - 2] + 1[(3)^3 - 3] + 4[(4)^3 - 4] + 1[(5)^3 - 5] + 2[(7)^3 - 7] + 1[(8)^3 - 8] + 1[(9)^3 - 9] + 1[(12)^3 - 12] + 1[(13)^3 - 13] + 1[(14)^3 - 14]}{100^3 - 100}\right]$$

$$= 0.991083$$

Zatem poprawiona wartość statystyki H wynosi:

$$H' = \frac{H}{C} = \frac{34.41}{0.991083} = 34.72$$

Hipotezę zerową odrzucamy, gdyż $H > \chi^2_{0.001,4} = 18.47$. Możemy zatem wnioskować, że liczba „łysinek" powstających w obecności każdego z badanych antybiotyków była istotnie różna, czyli testowane antybiotyki różnią się skutecznością niszczenia kultur bakterii.

Test mediany

Przykład 98

Dane z poprzedniego przykładu przeanalizujmy teraz korzystając z testu mediany dla więcej niż dwóch grup.

	antybiotyk 1	**antybiotyk 2**	**antybiotyk 3**	**antybiotyk 4**	**antybiotyk 5**
	2	2	6	1	4
	4	3	7	2	5
	4	4	7	3	5
	5	5	8	3	6
	6	5	8	3	7
	6	6	9	4	7
	6	6	10	5	7
	6	6	11	5	8
	7	7	11	5	9
	7	7	11	5	9
	7	7	12	5	10
	8	8	12	5	11
	8	9	12	6	11
	10	9	12	6	11
	11	9	12	6	12
	11	9	13	6	12
	13	10	13	7	12
	13	10	14	7	14
	14	11	14	7	15
	21	12	15	8	16
mediana	7	7	11.5	5	9.5
mediana całkowita	**7**				

	antybiotyk 1	**antybiotyk 2**	**antybiotyk 3**	**antybiotyk 4**	**antybiotyk 5**	***razem***
powyżej mediany	8	8	1	16	4	**37**
nie powyżej mediany	12	12	19	4	16	**63**
razem	**20**	**20**	**20**	**20**	**20**	**100**

Liczymy wartość statystyki chi^2:

$$\chi^2 = \sum \frac{\left(f_{obserwowane} - f_{oczekiwane}\right)^2}{f_{oczekiwane}} = 27.28$$

Hipotezę zerową odrzucamy, gdyż obliczona statystyka

$$\chi^2 > \chi^2_{0.001,41} = 9.49.$$

Testy do porównania dwóch zmiennych

Test kolejności par Wilcoxona

Przykład 99

Porównywano czas okluzji odzwierciedlający reaktywność płytek u pacjentów kardiochirurgicznych przed zabiegiem oraz 10 dni po zabiegu.

przed	po	przed	po
68	56	66	54
61	57	66	56
78	60	78	60
78	63	67	65
54	45	70	55
61	65	67	50
60	57	69	64
74	69	88	51
68	57	74	76
75	58	72	80
63	51	84	74
78	63	70	92
99	56	83	80
108	69	98	88
118	63		

Czy można uznać, że w późnym okresie po zabiegu występuje nadreaktywność płytek krwi, przejawiająca się skróconymi czasami okluzji?

Do obliczeń wykorzystamy test kolejności par Wilcoxona.

Dla niedużych liczebności test ten można przeprowadzić w prosty sposób w czterech etapach:

1. Wyrzucamy różnice wynoszące zero, zaś pozostałe porządkujemy w kolejności rosnącej, zaniedbując znak różnicy; w przypadku równych wartości różnic liczymy rangi wiązane;
2. Sumujemy rangi dla różnic dodatnich (T_+) i te dla różnic ujemnych (T_-);
3. Gdyby nie było różnic między zmiennymi, to sumy T_+ oraz T_- byłyby podobne; czym większa różnica między zmiennymi, tym większa byłaby różnica między T_+ i T_-; mniejsza z obliczonych sum stanowi statystykę testu T;
4. Porównujemy doświadczalną wartość T z wartością krytyczną w tablicach dla testu kolejności par Wilcoxona przy istotności α oraz liczebności N, oznaczającej liczbę różnic różnych od zera. Jeżeli wartość doświadczalna jest **mniejsza** od krytycznej tablicowej, to odrzucamy hipotezę zerową.

przed	*po*	*różnica*	*|różnica|*	*ranga*
67	65	**-2**	2	**1.5**
74	76	2	2	1.5
60	57	**-3**	3	**3.5**
83	80	**-3**	3	**3.5**
61	57	**-4**	4	**5.5**
61	65	4	4	5.5
69	62	**-7**	7	**7**
77	69	**-8**	8	**8.5**

72	80	8	8	8.5	
54	45	**-9**	9	**10**	
66	56	**-10**	10	**12**	
84	74	**-10**	10	**12**	
98	88	**-10**	10	**12**	
68	57	**-11**	11	**14**	
63	51	**-12**	12	**15.5**	
66	54	**-12**	12	**15.5**	
68	55	**-13**	13	**17**	
70	56	**-14**	14	**18**	
78	63	**-15**	15	**19**	
66	50	**-16**	16	**20.5**	
79	63	**-16**	16	**20.5**	
75	58	**-17**	17	**22**	
78	60	**-18**	18	**23**	
82	60	**-22**	22	**24.5**	
70	92	22	22	24.5	$T_- = 395$
88	51	**-37**	37	**26**	
108	69	**-39**	39	**27**	
99	56	**-43**	43	**28**	
118	63	**-55**	55	**29**	$T_+ = 40$

Wartość statystyki T to mniejsza spośród dwóch wartości T_+ i T_-: w naszym przypadku $T_+ = 40$.

Sprawdzamy, czy statystyka T (dla *n* równego liczbie analizowanych różnic), mniejsza spośród dwóch wartości T_+ i T_-, jest mniejsza niż można by oczekiwać przez czysty przypadek. Wynik jest istotny, jeżeli obliczona mniejsza spośród dwóch wartości T_+ i T_- jest **mniejsza** (!) niż wartość krytyczna. Przy $n = 29$ nasz wynik jest istotny na poziomie o wiele niższym niż 0.01 (leży „na prawo" od punktu krytycznego 110) dla testu jednostronnego.

Pamiętając, że w miarę wzrostu liczebności zmiennych, statystyki testów nieparametrycznych aproksymują do rozkładu normalnego, możemy także obliczyć wartość statystyki testu kolejności par Wilcoxona w alternatywny sposób. Hipoteza zerowa w tym teście zakłada, że jeśli nie ma różnic między zmiennymi, oczekiwana suma rang powinna być równo rozdzielona między dwie grupy: różnic ze znakiem (–) oraz różnic ze znakiem (+):

$$T_{oczekiwana} = \frac{n(n+1)}{2} * \frac{1}{2} = \frac{n(n+1)}{4} = \frac{29(29+1)}{2} * \frac{1}{2} = 217.5$$

Wartość statystyki testu wynosi:

$$z = \frac{T_{dośw} - T_{oczekiwane}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} = \frac{40 - 217.5}{\sqrt{\frac{29(29+1)(58+1)}{24}}} = -3.838$$

jeżeli za $T_{dośw}$ przyjmiemy $T_+ = 40$, lub $z = 3.838$, jeżeli za $T_{dośw}$ weźmiemy $T_- = 395$.

Taka wartość *z* odpowiada prawdopodobieństwu $\alpha < 0.0002$, możemy zatem odrzucić hipotezę zerową zakładającą brak różnicy między zmiennymi.

Test znaków

Przykład 100

Badano skuteczność antagonisty czynnika von Willebranda – kwasu aurinotrikarboksylowego (ATA) na skuteczność hamowania indukowanej przez ristocetynę aglutynacji płytek krwi. Stosując test znaków należy określić czy ATA znacząco hamuje aglutynację płytek krwi.

aglutynacja [%]			
bez ATA	**z ATA**	**różnica**	**znak**
7.8	16.8	9.0	–
23.8	15.8	–8.0	+
30.8	8.0	–22.8	+
27.3	44.9	17.6	–
75.0	28.2	–46.8	+
44.5	41.1	–3.4	+
30.4	33.8	3.3	–
51.3	56.2	4.9	–
69.9	76.0	6.2	–
30.6	83.2	52.6	–
51.8	48.1	–3.7	+
21.1	11.5	–9.6	+
54.5	74.3	19.9	–
21.7	31.4	9.6	–
56.3	14.5	–41.9	+

Znakiem (+) oznaczono przypadki, w których stwierdzono skuteczne hamowanie aglutynacji przez ATA, czyli dla których różnica [z ATA] – [bez ATA] była ujemna

Nasze hipotezy mają postać:

– H_0: nie ma istotnych różnic między pomiarami z ATA i bez ATA, czyli $p(+) = 0.50$
– H_A: są istotne różnice między pomiarami z ATA i bez ATA, czyli $p(+) \neq 0.50$

Czym bardziej proporcja $p(+)$ różni się od 0.50, tym większe jest prawdopodobieństwo, że różnice między zmiennymi nie są dziełem przypadku, lecz prawidłowością. W naszym przypadku udowodnienie, że $p(+)$ znacząco różni się od 0.50 będzie dowodem, że ATA znacząco hamuje aglutynację płytek indukowaną ristocetyną.

Dla prób o niewielkich liczebnościach (<10-15) posługujemy się logiką rachunku prawdopodobieństwa dla testu dwumianowego. Naszym zadaniem będzie obliczenie prawdopodobieństwa, że przynajmniej 7 z 15 porównań posiada znak (+), wiedząc, że losowe prawdopodobieństwo wystąpienia (+) wynosi 50%. Zauważmy, że gdyby wystąpienie znaku (+) wynikało jedynie z przypadku, a nie z tendencji, że wpływ ATA – jeśli jest istotny – zwiększa szansę pojawienia się znaku (+), to powinniśmy oczekiwać wystąpienia znaku (+) z prawdopodobieństwem 50% oraz znaku (–) także z prawdopodobieństwem 50%.

Prawdopodobieństwo, że przynajmniej 7 z 15 porównań posiada znak (+), będzie sumą prawdopodobieństw, że:

7 z 15 porównań posiada znak (+), (LUB)
8 z 15 porównań posiada znak (+), (LUB)
9 z 15 porównań posiada znak (+), (LUB)
10 z 15 porównań posiada znak (+), (LUB)
11 z 15 porównań posiada znak (+), (LUB)

12 z 15 porównań posiada znak (+), (LUB)
13 z 15 porównań posiada znak (+), (LUB)
14 z 15 porównań posiada znak (+), (LUB)
15 z 15 porównań posiada znak (+)

Czyli:

$$p(x) = \binom{n}{x} p^x q^{n-x} = \frac{n!}{x!(n-x)!}(p^x q^{n-x})$$

$$p(7) = \binom{15}{7}(0.50)^7 (0.50)^8 = \frac{15!}{7!(15-7)!}(0.50)^7 (0.50)^8$$

$$p(8) = \binom{15}{8}(0.50)^8 (0.50)^7 = \frac{15!}{8!(15-8)!}(0.50)^8 (0.50)^7$$

$$p(9) = \binom{15}{9}(0.50)^9 (0.50)^6 = \frac{15!}{9!(15-9)!}(0.50)^9 (0.50)^6$$

$$p(10) = \binom{15}{10}(0.50)^{10} (0.50)^5 = \frac{15!}{10!(15-10)!}(0.50)^{10} (0.50)^5$$

$$p(11) = \frac{15!}{11!(15-11)!}(0.50)^{11} (0.50)^4$$

$$p(12) = \frac{15!}{12!(15-12)!}(0.50)^{12} (0.50)^3$$

$$p(13) = \frac{15!}{13!(15-13)!}(0.50)^{13} (0.50)^2$$

$$p(14) = \frac{15!}{14!(15-14)!}(0.50)^{14} (0.50)^1$$

$$p(15) = \frac{15!}{14!(15-15)!}(0.50)^{15} (0.50)^0$$

x	$n-x$	n	$x!$	$(n-x)!$	$n!$	$x!(n-x)!$	$n!/[x!(n-x)!]$	p^x	q^{n-x}	$p^x q^{n-x}$	$p(x)$
7	8	15	5040	40320	1.31×10^{12}	203212800	6435	0.007813	0.003906	3.052×10^{-5}	0.196
8	7	15	40320	5040	1.31×10^{12}	203212800	6435	0.003906	0.007813	3.052×10^{-5}	0.196
9	6	15	362880	720	1.31×10^{12}	261273600	5005	0.001953	0.015625	3.052×10^{-5}	0.153
10	5	15	3628800	120	1.31×10^{12}	435456000	3003	0.000977	0.03125	3.05176×10^{-5}	0.0916
11	4	15	39916800	24	1.31×10^{12}	958003200	1365	0.000488	0.0625	3.05176×10^{-5}	0.0417
12	3	15	4.79×10^{8}	6	1.31×10^{12}	2.874×10^{9}	455	0.000244	0.125	3.05176×10^{-5}	0.0139
13	2	15	6.23×10^{9}	2	1.31×10^{12}	1.245×10^{10}	105	0.000122	0.25	3.05176×10^{-5}	0.0032
14	1	15	8.72×10^{10}	1	1.31×10^{12}	8.718×10^{10}	15	6.1×10^{-5}	0.5	3.05176×10^{-5}	0.00046
15	0	15	1.31×10^{12}	1	1.31×10^{12}	1.308×10^{12}	1	3.05×10^{-5}	1	3.05176×10^{-5}	3.1×10^{-5}
										Σ	**0.696**

p (>6 znaków (+)) = 0.696

Jest prawie 70% szans, że mniejsza aglutynacja płytek w obecności ATA (znak (+)) wystąpi przez czysty przypadek. Zatem nie ma podstaw do odrzucenia hipotezy zerowej.

Ponieważ liczność próby wynosi 15 przypadków, możemy także zastosować test proporcji z poprawką na ciągłość Yatesa:

$$z = \frac{\left|p - P_0\right| - \frac{1}{n}}{\sqrt{\frac{(P_0)(1-P_0)}{n}}} \qquad p = \frac{(+)}{(+)+(-)} = \frac{7}{15} = 0.4667 \quad P_0 = 0.5 \quad \text{i} \quad 1 - P_0 = 0.5 \quad n = 15$$

$$z = \frac{\left|p - P_0\right| - \frac{1}{n}}{\sqrt{\frac{(P_0)(1-P_0)}{n}}} = \frac{\left|0.4667 - 0.5\right| - \frac{1}{15}}{\sqrt{\frac{(0.50)(0.50)}{15}}} = \frac{0.0333 - 0.0667}{\sqrt{0.01667}} = -0.2582$$

Pole pod krzywą w punkcie $z = -0.26$ wynosi 0.3984, czyli prawdopodobieństwo, że wartość krytyczna znajdzie się powyżej (na prawo) wartości $z = -0.26$ będzie 1 – 0.3984 = 0.6016, czyli nie ma podstaw do odrzucenia hipotezy zerowej.

Test McNemara

Przykład 101

W grupie pacjentów z chorobą niedokrwienną serca (ChNS) zbadano wpływ leczenia hipolipemizującego przy użyciu ceriwastatyny (CRV) na wrażliwość płytek krwi na działanie aspiryny (ASA) określaną przy pomocy analizatora funkcji płytek PFA-100. Czy istnieje zależność między leczeniem CRV a wrażliwością płytek krwi na ASA?

	przed terapią CRV		
po terapii CRV	**„oporne" na ASA**	**wrażliwe na ASA**	***razem***
„oporne" na ASA	18	1	**19**
wrażliwe na ASA	73	89	**162**
razem	**91**	**90**	**181**

Ponieważ mamy dwie sparowane (pomiar przed i po terapii CRV) zmienne dyskretne (występuje oporność-brak oporności) z dwoma poziomami każda (oporne-wrażliwe, przed terapią-po terapii), stosujemy test McNemara. Aby wypełnić tabelę czteropolową liczymy:
- u ilu pacjentów, którzy byli „oporni" na ASA przed podaniem CRV oporność taka utrzymała się po zakończeniu terapii (*a*),
- u ilu pacjentów, którzy byli „oporni" na ASA przed podawaniem CRV oporności takiej nie stwierdzano po zakończeniu terapii (*c*),
- u ilu pacjentów, u których nie stwierdzano „oporności" na ASA przed podawaniem CRV nie wykryto „oporności" także po zakończeniu terapii (*d*), oraz
- u ilu pacjentów, u których nie stwierdzano „oporności" na ASA przed podawaniem CRV wykryto oznaki takiej „oporności" po zakończeniu terapii (*b*).

Statystykę testu obliczamy w następujący sposób:

$$\chi^2_{McNemara} = \frac{(b-c)^2}{b+c} = \frac{(3-37)^2}{3+37} = \frac{(-34)^2}{40} = 28.9$$

lub z uwzględnieniem poprawki na ciągłość Yatesa:

$$\chi^2_{McNemara} = \frac{(|b-c|-1)^2}{b+c} = \frac{(|3-37|-1)^2}{3+37} = 27.23$$

Ponieważ $\chi^2 = 27.23 > \chi^2_{0.05,1}$, odrzucamy hipotezę zerową mówiącą, że nie istnieje zależność między leczeniem CRV a wrażliwością na ASA.

Dla dużych liczebności możemy także policzyć statystykę testu proporcji, któremu jest równoważny test chi^2 McNemara ($\chi^2 = z^2$):

$$z = \frac{a-d}{\sqrt{a+d}} = \frac{16-89}{\sqrt{16+89}} = \frac{-73}{10.247} = -7.124\,.$$

Dla $z = -7.12$ z prawdopodobieństwem ponad 99.99% możemy odrzucić hipotezę zerową i stwierdzić, że jest istotna zależność między stosowaniem terapii CRV a wrażliwością płytek krwi na ASA.

Przykład 102

Grupę laborantów poproszono o wyrażenie swojej opinii na temat pipet wielokanałowych znanej firmy. Po wypełnieniu ankiety każdemu z respondentów powierzono pipetę, której dotyczyła ankieta, i poproszono o testowanie pipety przez okres miesiąca. Po tym okresie ponownie przeprowadzono ankietę z prośbą o wyrażenie opinii na temat pipety po raz drugi. Na podstawie uzyskanych wyników należy zbadać czy praca z testowaną pipetą przez okres miesiąca miała wpływ na opinie laborantów.

	przed testowaniem pipety		
po testowaniu pipety	**opinia pozytywna**	**opinia negatywna**	***razem***
opinia pozytywna	12	8	**20**
opinia negatywna	3	7	**10**
razem	**15**	**15**	**30**

Liczymy statystykę testu z uwzględnieniem poprawki na ciągłość Yatesa:

$$\chi^2_{McNemara} = \frac{(|b-c|-1)^2}{b+c} = \frac{(|8-3|-1)^2}{8+3} = \frac{(5-1)^2}{11} = \frac{16}{11} = 1.45$$

Ponieważ $\chi^2_{0.05,1} = 3.84 > \chi^2_{McNemara} = 1.45$, nie mamy podstaw aby odrzucić hipotezę zerową: testowanie pipety nie miało wpływu na opinię laborantów.

Test Friedmana

Przykład 103

Przy użyciu testu Friedmana należy sprawdzić, czy wpływ 4 różnych antagonistów płytek krwi na hamowanie adhezji płytek do kolagenu jest taki sam, czy różny.

	antagoniści			
bloki danych	**1**	**2**	**3**	**4**
1	7	5.3	4.9	8.8
rangi	***3***	***2***	***1***	***4***
2	9.9	5.7	7.6	8.9
rangi	***4***	***1***	***2***	***3***
3	8.5	4.7	5.5	8.1
rangi	***4***	***1***	***2***	***3***
4	5.1	3.5	2.8	3.3
rangi	***4***	***3***	***1***	***2***
5	10.3	7.7	8.4	9.1
rangi	***4***	***1***	***2***	***3***
suma rang	**19**	**8**	**8**	**15**

Każdy blok danych obejmuje pomiary wykonane u tego samego osobnika.

H_0: stopień hamowania adhezji płytek do kolagenu jest taki sam dla wszystkich badanych antagonistów

H_A: stopień hamowania adhezji płytek do kolagenu nie jest taki sam dla wszystkich badanych antagonistów

$k = 4$

$n = 5$

Statystykę testu liczymy jako:

$$\chi_r^2 = \frac{12}{nk(k+1)} \sum (R_j)^2 - 3n(k+1)$$

$$\chi_r^2 = \frac{12}{(5)(4)(4+1)} \left[19^2 + 8^2 + 8^2 + 15^2\right] - (3)(5)(4+1) = 85.68 - 75 = 10.68$$

Ponieważ $\chi^2 > \chi_{0.05,4,5} = 7.8$, (z tablic krytycznych wartości testu Friedmana), możemy odrzucić hipotezę zerową i stwierdzić, że stopień hamowania adhezji płytek do kolagenu nie jest taki sam dla wszystkich badanych antagonistów.

Test Cochrana

Przykład 104

20 ochotników uczestniczyło w badaniach wpływu 3 różnych blokerów receptorów płytkowych. Badania przeprowadzone w warunkach pozaustrojowych polegały na inkubowaniu próbek pełnej krwi z określonym preparatem przeciwpłytkowym oraz pomiarze czasu okluzji w kasetach z kolagenem i ADP. Podwyższenie czasu okluzji powyżej wartości uznanej za górną granicę zakresu normy określano jako obniżenie reaktywności płytek i przyporządkowano jej wartość 1. Jeżeli czas okluzji nie przekraczał tej górnej granicy normy, wynik opisywano jako 0. Czy trzy testowane preparaty różniły się skutecznością hamowania reaktywności płytek?

		preparat			
Lp	**A**	**B**	**C**	**R**	**R^2**
1	0	1	0	1	1
2	1	0	1	2	4
3	0	0	0	0	0
4	0	0	0	0	0
5	0	1	1	2	4
6	0	0	0	0	0
7	0	0	0	0	0
8	0	0	0	0	0
9	0	0	0	0	0
10	1	1	0	2	4
11	0	0	1	1	1
12	0	0	0	0	0
13	1	0	1	2	4
14	0	0	0	0	0
15	0	0	0	0	0
16	0	0	0	0	0
17	1	1	0	2	4
18	0	0	0	0	0
19	1	0	1	2	4
20	1	1	1	3	9
C =	6	5	6		
C^2 =	36	25	36		
			ΣR =	17	
	ΣC^2 =	97	ΣR^2 =		35

H_0: obniżanie reaktywności płytek nie zależy od rodzaju preparatu (blokera)
H_A: obniżanie reaktywności płytek zależy od rodzaju preparatu (blokera)
Wartość statystyki Q Cochrana wynosi:

$$Q = \frac{(k-1)\left[\left(k\Sigma C^2\right)-\left(\Sigma R\right)^2\right]}{k\left(\Sigma R\right)-\Sigma R^2} \qquad Q = \frac{(3-1)[3(97)-(17)^2}{3(17)-35} = \frac{4}{16} = 0.25$$

Ponieważ $Q < \chi^2_{0.05,2} = 5.99$, nie mamy podstaw do odrzucenia hipotezy zerowej.

Metody badania zależności – korelacja, regresja, wskaźniki zgodności między zmiennymi

Korelacja

Przykład 105

U 12 ochotników oceniano skuteczność blokera receptora dla fibrynogenu na hamowanie agregacji płytek krwi oraz obniżanie uwalniania selektyny P z ziarnistości a płytek.

dawca	agregacja	ranga	ekspresja selektyny P	ranga	d	d^2
1	3.2	3	2.7	2	1	1
2	4.7	4	6.3	6	–2	4
3	2.3	2	5.5	5	–3	9
4	1.9	1	1.6	1	0	0
5	8.8	8	10.2	10	–2	4
6	11.6	11	9.4	9	2	4
7	10.4	10	8.7	8	2	4
8	6.8	6	3.3	3	3	9
9	7.5	7	4.6	4	3	9
10	12.1	12	12.1	12	0	0
11	5.6	5	7.5	7	–2	4
12	9.9	9	11.8	11	–2	4
					Σd^2 =	**52**

Stosując metodę korelacji Spearmana należy obliczyć, czy istnieje zależność między tymi dwoma parametrami funkcji płytek krwi.

Współczynnik korelacji Spearmana obliczymy następująco:
1. Każdą zmienną osobno porządkujemy i nadajemy wartościom rangi, pamiętając o rangach wiązanych w przypadku identycznych wartości;
2. Obliczamy różnicę d dla każdej pary rang, liczymy kwadraty różnic i sumujemy je; nasze

$$\sum d^2 = 52;$$

3. Współczynnik korelacji Spearmana wynosi:

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6*52}{12(12^2-1)} = 1 - \frac{312}{1716} = 1 - 0.1818 = 0.818$$

4. Oceniamy istotność korelacji porównując obliczoną wartość z wartościami krytycznymi w tablicach dla danej liczby par; ponieważ obliczona przez nas wartość $r_s = 0.818 > r_s = 0.78$ dla 12 par wyników i $\alpha = 0.01$, możemy uznać, że zależność między zmiennymi nie jest przypadkowa.
5. Ponieważ liczba par jest większa od 10, możemy także policzyć istotność w sposób analogiczny jak dla współczynnika korelacji Pearsona:

$$t = r_s\sqrt{\left[\frac{n-2}{1-r_s^2}\right]} = 0.818\sqrt{\frac{12-2}{1-0.818^2}} = 0.818\sqrt{30.223} = 4.497$$

dla $d.f. = n - 2 = 12 - 2 = 10$

Ponieważ $t = 4.497 > t_{0.002.10} = 4.14$, możemy odrzucić hipotezę zerową mówiącą o niewystępowaniu zależności między tymi zmiennymi.

Przykład 106

Dla tych samych danych policzmy teraz współczynnik korelacji τ Kendalla.
Porządkujemy rangi jednej zmiennej w porządku rosnącym:

dawca	**4**	**3**	**1**	**2**	**11**	**8**	**9**	**5**	**12**	**7**	**6**	**10**
agregacja	1	2	3	4	5	6	7	8	9	10	11	12
selektyna P	1	5	2	6	7	3	4	10	11	8	9	12

Analizujemy kolejność rang drugiej zmiennej. Pierwsza para rang, 1-5 ma poprawną kolejność, gdyż 1 zawsze występuje przed 5; podobnie jest z parami: 1-2, 1-6, 1-7, itd. Każdej parze rang o prawidłowej kolejności przyporządkowujemy wartość (+1), o nieprawidłowej – wartość (–1). Dla rangi numer 1 całkowita punktacja wynosi 12 x (+1) = +12. Analogicznie, dla rangi 5 i następnych mamy:

											razem
5 –>	–1	+1	+1	–1	–1	+1	+1	+1	+1	+1	**+4**
	2 –>	+1	+1	+1	+1	+1	+1	+1	+1	+1	**+9**
		6 –>	+1	–1	–1	+1	+1	+1	+1	+1	**+4**
			7 –>	–1	–1	+1	+1	+1	+1	+1	**+3**
				3 –>	+1	+1	+1	+1	+1	+1	**+6**
					4 –>	+1	+1	+1	+1	+1	**+5**
						10 –>	+1	–1	–1	+1	**0**
							11 –>	–1	–1	+1	**–1**
								8 –>	+1	+1	**+2**
									9 –>	+1	**+1**
										+12 –>	**0**
										***razem*+44 (+12) = +56**	

Współczynnik τ Kendalla wynosi:

$$\tau = \frac{2S}{N(N-1)} = \frac{2(56)}{132} = 0.8485,$$

gdzie całkowita liczba porównywanych par zmiennych wynosi:

$$T = \frac{N(N-1)}{2} = \frac{12(11)}{2} = \frac{132}{2} = 66,$$

zaś S = + 56 oznacza sumę różnic punktacji zgodnych i niezgodnych kolejności dla porządku rosnącego rang.

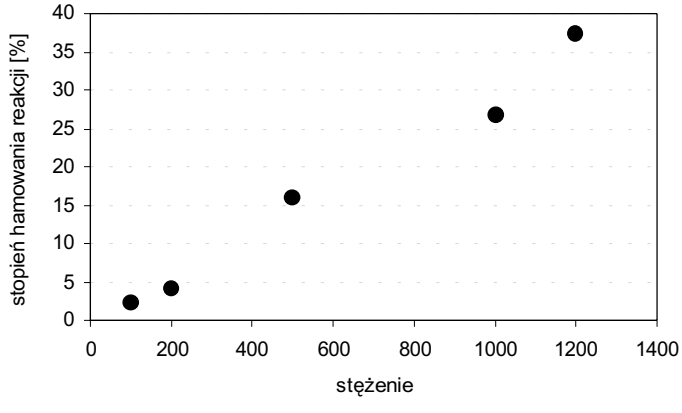
Regresja nieparametryczna

Przykład 107

Należy określić parametry równania regresji dla następujących danych:

	stężenie	hamowanie reakcji [%]
1	200	4.11
2	1200	37.42
3	500	15.88
4	1000	26.79
5	100	2.34

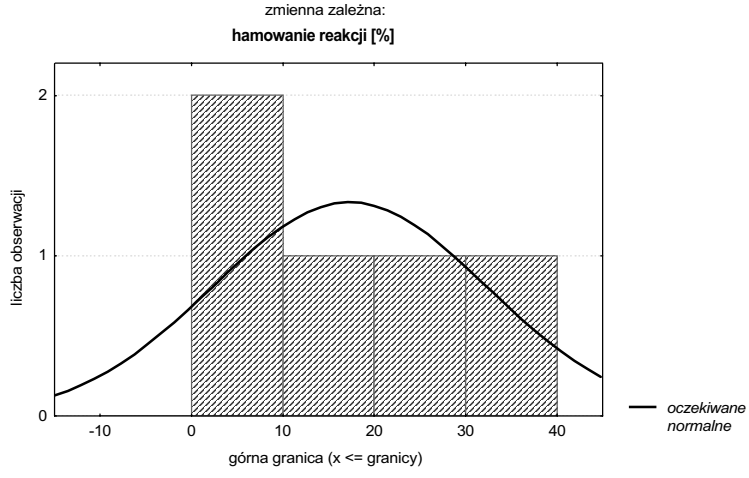
Sprawdzamy, czy zależność między danymi ma charakter liniowy; wykorzystujemy w tym celu metodę graficzną:



Sądząc po rozkładzie punktów na wykresie zależność ma charakter liniowy. Możemy w sposób roboczy – przed sprawdzeniem spełniania założenia o normalności zmiennej zależnej – policzyć współczynniki regresji. Wynoszą one:

	współczynniki	błąd standardowy	*t* Studenta	istotność (*p*)
przecięcie (*a*)	–1.05328	1.670921	–0.63036	0.573205
nachylenie (*b*)	0.030602	0.002257	13.5577	0.000868

Rozkład zmiennej zależnej odbiega od normalnego:



i dlatego zastosujemy metodę regresji nieparametrycznej – tzw. „niezupełną" metodę Theila, do wyznaczenia współczynników równania regresji.

Podobnie jak w przypadku wszystkich procedur nieparametrycznych, dane porządkujemy rosnąco pod względem zmiennej niezależnej *x*:

	stężenie	hamowanie reakcji [%]
1	100	2.34
2	200	4.11
3	500	15.88
4	1000	26.79
5	1200	37.42

Metoda Theila wymaga parzystej liczby par zmiennych, dlatego też środkową wartość (medianę) usuwamy.

Obliczamy wartość współczynnika *b*:

$$b_{ij} = \frac{(y_j - y_i)}{(x_j - x_i)}$$

dla dwóch par punktów:

$$b_{14} = \frac{(26.79 - 2.34)}{(1000 - 100)} = 0.0272 \qquad b_{25} = \frac{(37.42 - 4.11)}{(1200 - 200)} = 0.0333$$

Mediana tych dwóch wartości będzie naszym współczynnikiem $b = 0.03025$.
Dla każdej pary wartości x, y wyznaczamy współczynnik a:

$$a = y - bx$$

$$a_1 = 2.34 - (0.03025) * (100) = 2.34 - 3.025 = -0.685$$

$$a_2 = 4.11 - (0.03025) * (200) = 4.11 - 6.05 = -1.94$$

$$a_4 = 26.79 - (0.03025) * (1000) = 26.79 - 30.25 = -3.46$$

$$a_5 = 37.42 - (0.03025) * (1200) = 37.42 - 36.3 = 1.12$$

Współczynnikiem a linii regresji będzie wartość mediany dla 4 oszacowanych wartości, czyli:

–3.46 –1.94 –0.685 1.12

$$\frac{(-1.94) + (-0.685)}{2} = -1.3125$$

Estymowane równanie ma postać:

$$y = 0.03025x - 1.3125$$

Współczynniki zgodności dla zmiennych nominalnych

Przykład 108

Badano skuteczność nowego leku trombolitycznego w reperfuzji u nosicieli różnych wariantów polimorfizmu płytkowego $Pl^{A1/A2}$, u których wystąpił drugi zawał mięśnia sercowego:

	udana reperfuzja	**nieudana reperfuzja**	*razem*
$Pl^{A2}(-)$	84	16	**100**
$Pl^{A2}(+)$	64	36	**100**
razem	**148**	**52**	**200**

Czy na podstawie tych wstępnych obserwacji można powiedzieć, że skuteczność reperfuzji jest podobna u nosicieli obu wariantów polimorfizmu?

Możemy to pytanie postawić także w inny sposób: czy występowanie prozakrzepowego polimorfizmu $Pl^{A2}(+)$ koreluje z występowaniem nieudanej reperfuzji po zastosowaniu nowego leku trombolitycznego ?

Ponieważ mamy do czynienia z danymi nominalnymi (binarne układy: jest-nie ma, występuje-nie występuje), do zbadania korelacji posłużymy się wskaźnikami zgodności dla zmiennych nominalnych.

Wyznaczmy współczynnik korelacji czteropolowej ϕ Cramera według równania:

$$\phi_2 = \frac{(a)(d)-(b)(c)}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

gdzie

a	*b*	*a+b*
c	*d*	*c+d*
a+c	*b+d*	*n*

Współczynnik ϕ_2 wynosi:

$$\phi_2 = \frac{(64)(16)-(36)(84)}{\sqrt{(84+16)(64+36)(84+64)(16+36)}} = \frac{1024-3024}{\sqrt{(100)(100)(148)(52)}} = \frac{-2000}{877268} = -0.228$$

Niska wartość współczynnika ϕ Cramera wskazuje na niewielką zgodność i niewielkie podobieństwo pod względem skuteczności badanego leku u poddanych obserwacji nosicieli różnych wariantów polimorfizmu $Pl^{A1/A2}$: nosiciele wariantu Pl^{A2}(+) odpowiadają gorzej na testowany lek trombolityczny (ujemna wartość współczynnika).

Dla tych samych danych policzmy także wartości współczynnika Yule'a oraz współczynnika Ivesa-Gibbonsa. Uzyskujemy:

$$Q = \frac{(a)(d)-(b)(c)}{(a)(d)+(b)(c)} = \frac{(64)(16)-(36)(84)}{(64)(16)+(36)(84)} = \frac{1024-3024}{1024+3024} = -0.494,$$

oraz

$$r_n = \frac{(a+d)-(b+c)}{(a+d)+(b+c)} = \frac{(64+16)-(36+84)}{(64+16)+(36+84)} = \frac{80-120}{80+120} = -0.200$$

Zauważmy, że spośród porównanych wskaźników najmniejszą mocą odznacza się współczynnik Yule'a (zawyża rzeczywistą miarę zgodności), zaś największą – współczynnik Ivesa-Gibbonsa (zaniża miarę zgodności). Współczynnik ϕ Cramera zajmuje pozycję pośrednią.

Przykład 109

W grupie pacjentów poddanych zabiegowi kardiochirurgicznemu wszczepienia pomostów aortalno-wieńcowych oceniano częstość wystąpienia reokluzji naczyń oraz rejestrowano współwystępowanie dwóch prozakrzepowych alleli polimorfizmów glikoprotein płytek krwi: ^{807}T glikoproteiny Ia receptora dla kolagenu oraz Pl^{A2} glikoproteiny IIIa receptora dla

fibrynogenu. Należy ocenić czy współwystępowanie tych dwóch polimorfizmów w przebadanej grupie 32 pacjentów ma związek z wyższą częstością pozabiegowej reokluzji naczyń.

pacjent	**wystąpienie reokluzji**	**współwystępowanie alleli ^{807}T i Pl^{A2}**
1	–	+
2	+	+
3	–	–
4	–	+
5	–	–
6	+	+
7	+	+
8	+	–
9	–	+
10	+	+
11	+	+
12	–	–
13	+	+
14	–	–
15	+	+
16	–	–
17	–	+
18	+	+
19	–	–
20	–	–
21	+	+
22	+	–
23	–	–
24	–	–
25	+	+
26	–	+
27	–	–
28	+	+
29	–	–
30	+	–
31	–	–
32	+	+

Nasze pytanie możemy także sformułować następująco: czy występuje zależność statystyczna między wystąpieniem reokluzji a współwystępowaniem tych dwóch polimorfizmów?
Rozkład obserwacji w tabeli 2 x 2 wygląda następująco:

współwystępowanie alleli ^{807}T i Pl^{A2}

reokluzja	***jest***	***brak***	***razem***
jest	12	2	**14**
brak	5	13	**18**
razem	**17**	**15**	**32**

Współczynnik ϕ_2 Cramera wynosi:

$$\phi_2 = \frac{(a)(d)-(b)(c)}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

$$\phi_2 = \frac{(12)(13)-(2)(5)}{\sqrt{(12+2)(5+13)(12+5)(2+5)}} = \frac{156-10}{\sqrt{(14)(18)(17)(7)}} = \frac{146}{253.5} = 0.576$$

Wysoka wartość tego współczynnika oznacza, że istnieje zależność między współwystępowaniem alleli ^{807}T i Pl^{A2} u chorego a wystąpieniem reokluzji w następstwie zabiegu kardiochirurgicznego. Mogłoby to sugerować, że łączne występowanie obu prozakrzepowych alleli u tego samego pacjenta wiąże się ze zwiększonym ryzykiem epizodu reokluzji naczyń po zabiegu kardiochirurgicznym u tej osoby.

Dla danych z powyższej tabeli policzmy również wartości dwóch innych wskaźników: współczynnika Yule'a oraz współczynnika Ivesa-Gibbonsa.
Uzyskujemy wartości:

$$Q = \frac{(a)(d)-(b)(c)}{(a)(d)+(b)(c)} = \frac{(12)(13)-(2)(5)}{(12)(13)+(2)(5)} = \frac{156-10}{156+10} = 0.880$$

$$r_n = \frac{(a+d)-(b+c)}{(a+d)+(b+c)} = \frac{(12+13)-(2+5)}{(12+13)+(2+5)} = \frac{25-7}{25+7} = 0.563$$

Tak, jak w poprzednim przykładzie, współczynnik Ivesa-Gibbons przyjmuje najniższą wartość, zaś współczynnik Yule'a – najwyższą.

Ma to istotne implikacje w odniesieniu do wniosków płynących z takich obliczeń. Przyjmijmy hipotetycznie, że obserwowane przez nas częstości wariantów w tabeli 2 x 2 wynosiłyby:

współwystępowanie alleli ^{807}T i Pl^{A2}

reokluzja	***jest***	***brak***	***razem***
jest	12	2	**14**
brak	13	5	**18**
razem	**25**	**7**	**32**

Ewidentny brak zależności w tym przypadku (oba allele współwystępują w przybliżeniu jednakowo często wśród pacjentów, u których zaobserwowano reokluzję, jak i wśród tych, u których jej nie było) sprawia, że wartości badanych wskaźników są niższe:

$$\phi_2 = 0.162, \qquad Q = 0.395 \qquad \text{oraz} \qquad r_n = 0.063$$

Zwróćmy uwagę, że współczynnik Yule'a nadal przyjmuje dość wysoką wartość. Gdybyśmy mieli jedynie na podstawie tego jednego wskaźnika wypowiedzieć się o występowaniu bądź niewystępowaniu rzeczywistej zależności, trudno byłoby nam zdecydować, czy jego wartość jest już za niska, aby wskazywać na niewystępowanie zależności, czy też jeszcze wystarczająco wysoka, aby wskazywać na jej występowanie. Najlepszą dyskryminację występowania/braku zależności obserwujemy w przypadku współczynnika Ivesa-Gibbonsa: jest on „przekonująco" wysoki tam, gdzie zależność występuje, oraz „jawnie" niski w przypadku gdy jej brak.

Przykład 110

U mniej „podatnych" pacjentów (tych z polimorfizmem $Pl^{A2}(+)$) sprawdzano wpływ czasu trwania terapii trombolitycznej na skuteczność reperfuzji:

	udana reperfuzja	**nieudana reperfuzja**	***razem***
3 dni	13	17	**30**
10 dni	19	16	**35**
21 dni	21	14	**35**
razem	**53**	**47**	**100**

Czy możemy powiedzieć, że skuteczność leku trombolitycznego zależy od czasu trwania terapii?

Związek między zmiennymi dyskretnymi ocenimy stosując współczynnik zgodności C, obliczany na podstawie wartości statystyki testu chi^2 dla tablicy wielopolowej *c* x *r*:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

Dla sześciopolowej tabeli z dwoma kolumnami oraz trzema rzędami wartość statystyki testu chi^2 (nie skorygowana na ciągłość) wynosi:

$$\chi^2 = 33.2473 \qquad \text{oraz} \qquad C = \sqrt{\frac{33.2473}{33.2473 + 100}} = \sqrt{0.2495} = 0.4995$$

Ponieważ $\chi^2 = 33.2473 > \chi^2_{0.001,2}$, możemy odrzucić hipotezę zerową (nie ma zależności między czasem trwania terapii a skutecznością leku tromobolitycznego) i przyjąć hipotezę alternatywną stwierdzającą, że istnieje istotna zależność między czasem trwania terapii a skutecznością leku.

Badanie dopasowania rozkładu

Test zgodności chi^2 do badania dopasowania rozkładu

Przykład 111

Wykorzystując test zgodności chi^2 należy zbadać czy rozkład liczebności białych (17) i czerwonych (83) kwiatów groszku pachnącego w drugim pokoleniu krzyżówki osobnika o kwiatach czerwonych z osobnikiem o kwiatach białych jest zgodny z pierwszym prawem Grzegorza Mendla.

Przypomnijmy, że zgodnie z prawem Mendla oczekiwalibyśmy, aby stosunek kwiatów czerwonych do białych w drugim pokoleniu był jak 3:1, czyli wśród 100 zbadanych osobników groszku powinniśmy wykryć 75 osobników o kwiatach czerwonych i 25 osobników o kwiatach białych.

	czerwone	**białe**	***n***
$f_{obserwowane}$	83	17	100
$f_{oczekiwane}$	75	25	

Liczba stopni swobody wynosi $\nu = k - 1 = 2 - 1 = 1$.
a wartość statystyki:

$$\chi^2 = \sum \frac{\left(f_{obserwowane} - f_{oczekiwane}\right)^2}{f_{oczekiwane}} = \frac{(83-75)^2}{75} + \frac{(17-25)}{25} = \frac{8^2}{75} + \frac{8^2}{25} = \frac{64}{75} + \frac{64}{25} = 3.413$$

Ponieważ $\chi^2_{0.05,1} = 3.84$, nie ma podstaw do odrzucenia hipotezy zerowej mówiącej, że rozkład liczebności groszku o białych i czerwonych kwiatach jest zgodny z pierwszym prawem Mendla.

Przykład 112

Przy użyciu testu zgodności chi^2 należy sprawdzić czy rozkłady stężenia inhibitora tkankowego aktywatora plazminogenu (PAI-1), fibrynogenu (Fg) i całkowitego cholesterolu (TC) są normalne (dane źródłowe w arkuszu ***roz-normalne.xls***).

Aby udowodnić, że rozkład zmiennej posiada charakterystykę rozkładu normalnego, powinniśmy wykazać, że poszczególne wartości zmiennej pojawiają się z określonym prawdopodobieństwem, czyli odpowiadają polom powierzchni pod krzywą obliczonym na podstawie funkcji gęstości rozkładu normalnego. Pamiętamy, że takie przedziały pola powierzchni pod krzywą rozkładu (odpowiadające prawdopodobieństwom, że zmienna przyjmuje wartości z pewnego zakresu) możemy oszacować znając wartość średnią i wartość odchylenia zgodnie ze wzorem liczenia wartości krytycznych rozkładu normalnego:

$$z = \frac{x_i - \mu}{\sigma}$$

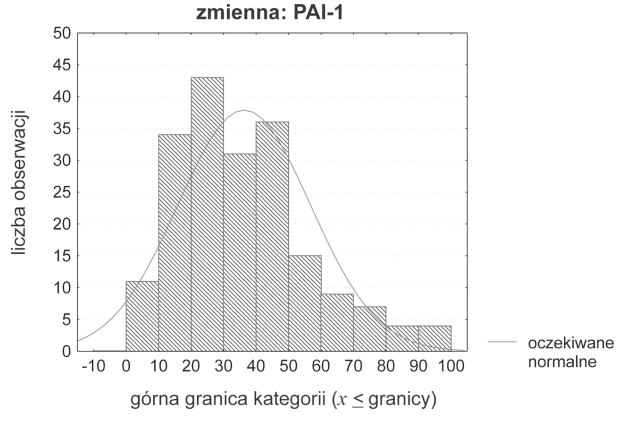
W rzeczywistości jednak nigdy nie znamy prawdziwych μ i σ, możemy jedynie z pewnym przybliżeniem oszacować wartość krytyczną na podstawie wartości średniej ($\bar{x}$) i odchylenia standardowego (s) z badanej próby i zapisać:

$$z = \frac{x_i - \bar{x}}{s}$$

Tak obliczoną wartość z nazywamy ***probitem*** (wartością normalną).

Przy weryfikacji normalności rozkładu za pomocą testu zgodności chi^2 powinniśmy:
- posortować dane w kolejności rosnącej;
- pogrupować wszystkie dane tworząc robocze kategorie (przedziały o stałej szerokości/ zakresie), pamiętając o tym, aby liczba kategorii nie była za mała, a średnia liczność obserwacji w każdej kategorii nie była niższa niż 5;
- przyporządkować każdej kategorii liczbę obserwacji dla danego zakresu zmiennej;
- policzyć częstości oczekiwane dla każdej kategorii;
- policzyć wartość statystyki chi^2 i porównać ją z wartością tablicową dla liczby stopni swobody $v = k - l - 1$, gdzie k oznacza liczbę kategorii (klas), a l liczbę parametrów rozkładu (dla rozkładu normalnego dwa: średnia i odchylenie standardowe); hipotezę zerową odrzucamy, gdy $\chi^2_{dośw} > \chi^2_{\alpha,v}$.

Graficzna prezentacja rozkładu tej zmiennej wskazuje, że jej rozkład jest wyraźnie prawoskośny:



Zauważmy, że jeśli liczności w poszczególnych klasach (kategoriach) przedstawimy jako składowe całkowitej liczebności próby, to ich suma (***skumulowana wartość*** poszczególnych składników) będzie reprezentowała całe pole pod krzywą. Na podstawie takich skumulowanych wartości dla każdego przedziału (kategorii) możemy policzyć wartość pola dla każdej kategorii (zakresu wartości) jako:

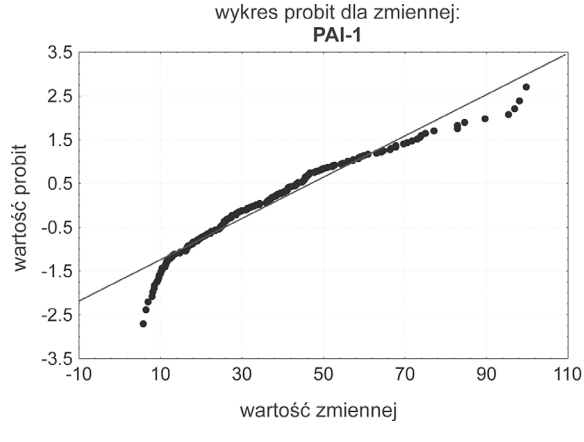
pole dla klasy (k) = pole skumulowane dla klasy (k) – pole skumulowane dla klasy (k-1)

zakres	liczność kategorii	% ogółu przypadków	% skumulowany ogółu przypadków
0–	11	5.67	5.67
10–	34	17.53	23.20
20–	43	22.16	45.36
30–	31	15.98	61.34
40–	36	18.56	79.90
50–	15	7.73	87.63
60–	9	4.64	92.27
70–	7	3.61	95.88
80–	4	2.06	97.94
90–	4	2.06	100.00
Σ	**194**	**100.00**	

Na przykład, pole dla klasy (40 –) obliczylibyśmy jako: 79.90 – 61.34 = 18.56.
Parametry zmiennej PAI-1 wynoszą:

$$\bar{x} = 36.33, \quad s = 20.42, \quad \text{wartość minimalna} = 5.81, \quad \text{wartość maksymalna} = 99.72$$

Obliczając wartości z dla poszczególnych zakresów (kategorii) podstawiamy do równania wartość górnej granicy każdego zakresu. Prawdopodobieństwa dla tak policzonych wartości krytycznych dla górnej granicy przedziału to po prostu wartości dystrybuanty rozkładu normalnego w punktach x_i wyznaczających górną granicę przedziału. Prawdopodobieństwa te znajdujemy w tablicach pola powierzchni pod krzywą rozkładu normalnego. Na przykład, jeżeli $z = -2.0$ to prawdopodobieństwo, że zmienna znajdzie się w obszarze pola pod krzywą, dla którego wartość –2.0 jest górną granicą, wynosi 0.02275. Jak to obliczamy? W tablicy pola pod krzywą znajdziemy prawdopodobieństwa rozkładu normalnego dla odpowiednich wartości statystyki z tego, że zmienna znajdzie się na prawo od z. Tablice podają jednak jedynie wartości prawdopodobieństw w zakresie 0 – 0.5 dla dodatnich z. Rozkład normalny jest jednak rozkładem idealnie symetrycznym, a osią tej symetrii jest średnia równa 0. Zatem na podstawie podawanych w tablicach wartości prawdopodobieństw dla dodatnich z można obliczać prawdopodobieństwa (pola pod krzywą) dla ujemnych z różnicy: 0.5 – odczytane prawdopodobieństwo dla dodatniej wartości $z + 0.5$. I tak, dla $z = +2.0$ wartość tablicowa wynosi 0.02275, czyli dla $z = -2.0$ wartość ta będzie wynosić $0.5 - 0.02275 + 0.5 = 0.97725$ (czyli $1 - 0.02275 = 0.97725$). Tablicowe wartości to jednak pola leżące na prawo od zadanej wartości z, nas natomiast interesują wartości na lewo od z, tzn. takie dla których z stanowi kres górny (górną granicę przedziału/kategorii), czyli wartości dystrybuanty rozkładu w poszczególnych punktach z. Jeżeli więc pole leżące na prawo od $z = -2.0$ wynosi 0.97725 całości pola pod krzywą, to część pola leżąca na lewo i kończąca się w punkcie $z = -2.0$ będzie wynosiła 0.02275. Zauważmy więc, że poszukiwana przez nas wartość dystrybuanty dla ujemnych z odpowiada wartości odczytanej z tablic dla dodatnich z (z uwagi na symetryczność rozkładu), natomiast szukana przez nas wartość dla dodatnich z będzie wynosić: 1.0 – wartość tablicowa. Rozkład wartości probit policzonych dla zmiennej PAI-1 wygląda tak:



Wykres probitowy wskazuje, że przy wysokich wartościach zmiennej, a w szczególności przy jej najniższych wartościach, nie jest zachowana liniowość zależności wartości probit od wartości zmiennej. Taka charakterystyczna nieliniowość jest typowa dla zmiennych o rozkładach prawoskośnych. Widzimy, że ta graficzna „metoda probitowa" może być dla nas użytecznym narzędziem do wstępnej oceny, czy zmienna posiada rozkład normalny lub jak dalece jest on naruszony. W zastosowaniu do zmiennej PAI-1 metoda ta pozwoliła nam się dowiedzieć, że rozkład tej zmiennej odbiega od rozkładu normalnego. Weryfikacji rachunkowej takiego przypuszczenia dostarcza nam test zgodności chi^2.

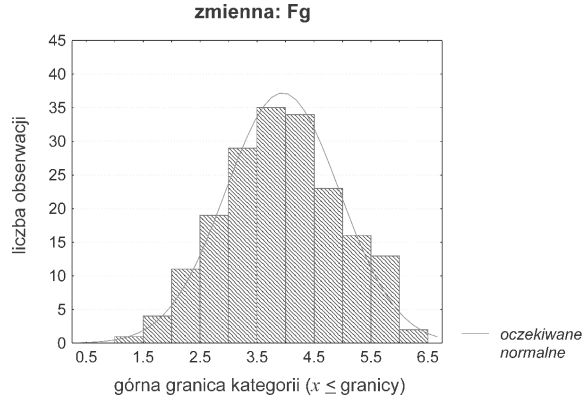
Na podstawie znajomości dystrybuant dla górnych granic (d_i) każdego przedziału obliczamy pole pod krzywą odpowiadające określonemu zakresowi jako: $d_i - d_{i-1}$.

zakres przedziału	**liczność klasy**	z	**dystrybuanta**	**liczności oczekiwane**	**liczności oczekiwane dla zakresu**	$(f_{obs}-f_{ocz})$	$(f_{obs}-f_{ocz})^2/f_{ocz}$
$-10 < x \le 0$	0	–1.78	0.03750	7.27500	7.27500	7.27500	7.275
$0 < x \le 10$	11	–1.29	0.09850	19.10900	11.83400	0.83400	0.058
$10 < x \le 20$	34	–0.80	0.21190	41.10860	21.99960	–12.00040	6.546
$20 < x \le 30$	43	–0.31	0.37830	73.39020	32.28160	–10.71840	3.559
$30 < x \le 40$	31	0.18	0.57140	110.85160	37.46140	6.461400	1.114
$40 < x \le 50$	36	0.67	0.74860	145.22840	34.37680	–1.62320	0.077
$50 < x \le 60$	15	1.16	0.87700	170.13800	24.90960	9.90960	3.942
$60 < x \le 70$	9	1.65	0.95050	184.39700	14.25900	5.25900	1.939
$70 < x \le 80$	7	2.14	0.98382	190.86108	6.46408	–0.53592	0.044
$80 < x \le 90$	4	2.62	0.99573	193.17162	2.31054	–1.68946	1.235
$90 < x \le 100$	4	3.12	0.99910	193.82540	0.65378	–3.34622	17.127
	194					**Σ =**	**42.92**

Ponieważ $\chi^2 = 42.92 > \chi^2_{0.001,8} = 26.13$, możemy z prawdopodobieństwem 99.9% odrzucić hipotezę zerową mówiącą o tym, że rozkład zmiennej PAI-1 jest normalny.
Analogicznie sprawdzamy rozkład zmiennej Fg:

$\bar{x} = 4.216$, $s = 1.305$, wartość minimalna = 1.3, wartość maksymalna = 8.0

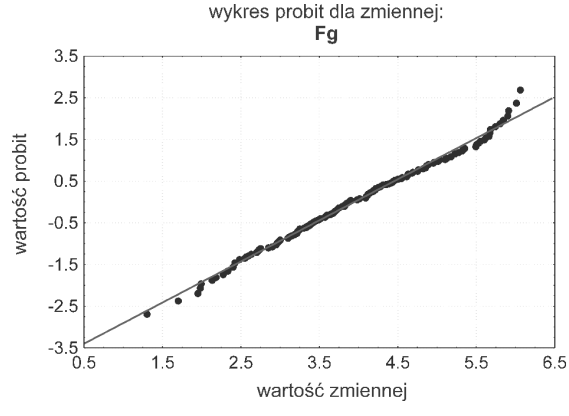
Wykres histogramowy rozkładu zmiennej: Fg wskazuje, że rozkład ten jest symetryczny:



Przeprowadzamy obliczenia statystyki chi^2:

zakres przedziału	liczność klasy	z	dystrybuanta	liczności oczekiwane	liczności oczekiwane dla zakresu	$(f_{obs}-f_{ocz})$	$(f_{obs}-f_{ocz})^2/f_{ocz}$
0.50 < x ≤ 1.0	0	–2.94	0.00676	1.26	1.264	1.264	1.264
1.0 < x ≤ 1.5	1	–2.44	0.01876	3.51	2.244	1.244	0.690
1.5 < x ≤ 2.0	4	–1.94	0.0446	8.34	4.832	0.832	0.143
2.0 < x ≤ 2.5	11	–1.44	0.0934	17.47	9.126	–1.874	0.385
2.5 < x ≤ 3.0	19	–0.95	0.1762	32.95	15.484	–3.516	0.799
3.0 < x ≤ 3.5	29	–0.45	0.2912	54.45	21.505	–7.495	2.612
3.5 < x ≤ 4.0	35	0.05	0.4325	80.88	26.423	–8.577	2.784
4.0 < x ≤ 4.5	34	0.55	0.5871	109.79	28.910	–5.090	0.896
4.5 < x ≤ 5.0	23	1.05	0.7257	135.71	25.918	2.918	0.329
5.0 < x ≤ 5.5	16	1.55	0.8365	156.43	20.720	4.720	1.075
5.5 < x ≤ 6.0	13	2.05	0.9147	171.05	14.623	1.623	0.180
6.0 < x ≤ 6.5	2	2.55	0.9599	179.50	8.452	6.452	4.926
	187					**Σ =**	**16.082**

Punkty na wykresie *wartości probit vs. wartości zmiennej* rozkładają się na linii prostej, co charakteryzuje rozkłady symetryczne, takie jak na przykład rozkład normalny:



Widzimy, że takiej liniowości nie spełniają w pełni najwyższe wartości zmiennej. Decyduje to o tym, że obliczona wartość χ^2 jest bardzo bliska wartości granicznej (16.92), zatem niewiele nas dzieli, abyśmy mogli odrzucić H_0. Innymi słowy, gdybyśmy odrzucili H_0, to nie popełnilibyśmy aż tak dużego błędu I rodzaju (α). Ponieważ jednak $\chi^2 = 16.08 < \chi^2_{0.05,9} = 16.92$, nie mamy podstaw do odrzucenia hipotezy zerowej zakładającej normalność rozkładu zmiennej Fg.

Na koniec, sprawdzamy rozkład zmiennej TC:

$\bar{x} = 230.761$, $s = 45.394$, wartość minimalna = 113, wartość maksymalna = 372

zakres przedziału	liczność klasy	z	dystrybuanta	liczności oczekiwane	liczności oczekiwane dla zakresu	$(f_{obs}-f_{ocz})$	$(f_{obs}-f_{ocz})^2/f_{ocz}$
$80 < x \leq 100$	0	–2.88	0.0009	0.18	0.185	0.185	0.185
$100 < x \leq 120$	1	–2.44	0.00427	0.88	0.691	–0.309	0.138
$120 < x \leq 140$	6	–2.00	0.01659	3.40	2.526	–3.474	4.780
$140 < x \leq 160$	5	–1.56	0.0516	10.58	7.177	2.177	0.660
$160 < x \leq 180$	12	–1.12	0.1292	26.49	15.908	3.908	0.960
$180 < x \leq 200$	30	–0.68	0.2643	54.18	27.696	–2.305	0.192
$200 < x \leq 220$	34	–0.24	0.4483	91.90	37.720	3.720	0.367
$220 < x \leq 240$	34	0.20	0.6406	131.32	39.422	5.422	0.746
$240 < x \leq 260$	34	0.64	0.8051	165.05	33.723	–0.278	0.002
$260 < x \leq 280$	18	1.08	0.9131	187.19	22.140	4.140	0.774
$280 < x \leq 300$	15	1.53	0.9686	198.56	11.378	–3.622	1.153
$300 < x \leq 320$	11	1.97	0.99086	203.13	4.563	–6.437	9.079
$320 < x \leq 340$	3	2.41	0.99788	204.57	1.439	–1.561	1.693
$340 < x \leq 360$	1	2.85	0.9996	204.92	0.353	–0.647	1.189
$360 < x \leq 380$	1	3.29	0.99994	204.99	0.070	–0.930	12.417
$380 < x \leq 400$	0	3.73	0.99999	205.00	0.010	0.010	0.010
	205					**Σ =**	**34.34497**

Ponieważ $\chi^2 = 34.345 > \chi^2_{0.005,13} = 29.82$, możemy z prawdopodobieństwem 99.5% odrzucić hipotezę zerową mówiącą o tym, że rozkład zmiennej TC jest normalny.

Przykład 113

Zważono 148 uczniów szkoły podstawowej i masę każdego dziecka zanotowano z dokładnością do jednego miejsca po przecinku. Uzyskane wyniki przedstawiono w tabeli:

cyfra w miejscu dziesiętnym	częstości obserwowane (f_{obs})	częstości oczekiwane (f_{oczek})	$(f_{obs}-f_{oczek})^2/f_{oczek}$
0	18	14.8	0.691892
1	14	14.8	0.043243
2	21	14.8	2.597297
3	12	14.8	0.52973
4	15	14.8	0.002703
5	19	14.8	1.191892
6	12	14.8	0.52973
7	17	14.8	0.327027
8	11	14.8	0.975676
9	9	14.8	2.272973
Σ	**148**		**9.16**

Czy rozkład wyników jest zgodny z rozkładem losowym?

Gdyby waga nie przekłamywała wartości pomiaru i nie rejestrowała wyników albo zaokrąglonych do całego albo do połówki kilograma, wówczas powinny wystąpić równe liczby powtórzeń z jedynką w miejscu dziesiętnym, z dwójką, trójką, czwórką, itd. Jeżeli wykonaliśmy 148 pomiarów ważenia, to każda z dziesięciu cyfr miejsca dziesiętnego powinna się pojawić średnio 14.8 razy. Taki teoretyczny rozkład losowy musimy porównać z rozkładem obserwowanym (*zobacz tabela powyżej*).

Widzimy, że obliczona wartość statystyki χ^2 = 9.16. Dla 9 stopni swobody (ponieważ mamy 10 różnych wariantów cyfr w miejscu dziesiętnym i nie musimy ustalać żadnego parametru naszego losowego rozkładu, zatem *d.f.* = 10 – 0 – 1 = 9), taka wartość statystyki wypada między wartościami prawdopodobieństwa 25% i 50%, czyli obserwowane częstości wyników ważenia pokrywają się z rozkładem teoretycznym.

Przykład 114

Badano liczbę komórek zarodźca malarycznego przypadających na jedną krwinkę czerwoną. Czy na podstawie zaobserwowanych liczebności krwinek z więcej niż jedną komórką pasożyta, możemy stwierdzić, że podwyższone częstości zakażeń wielokrotnych w przypadku zarodźca malarycznego nie są dziełem przypadku?

liczba komórek pasożyta na krwinkę	częstości obserwowane	częstości oczekiwane wg rozkładu Poissona	$(f_{obs}-f_{oczek})^2/f_{oczek}$
0	54168	53582.26	6.40
1	9257	10350.44	115.51
2	1460	999.69	211.95
3	81	64.37	4.30
4	34	3.11	306.98
>5	0	0.12	0.12
Σ	**65000**	**65000.00**	**645.26**

Gdyby infekcje wielokrotne były dziełem przypadku, to liczba komórek pasożyta przypadająca na pojedynczą krwinkę podlegałaby rozkładowi Poissona. Aby to sprawdzić, należałoby porównać częstości obserwowane z częstościami oczekiwanymi oszacowanymi na podstawie funkcji prawdopodobieństwa dla rozkładu Poissona. W tym celu policzmy średnią liczbę komórek zarodźca przypadających na jedną krwinkę:

$$\mu = \frac{0*54168 + 1*9257 + 2*1460 + 3*81 + 4*34 + 5*0}{65000} = \frac{12556}{65000} = 0.193$$

Prawdopodobieństwa znalezienia 0, 1, 2, 3, 4 i >5 komórek zarodźca w krwince liczymy z funkcji gęstości rozkładu Poissona:

$$P(x) = \frac{e^{-\mu}\mu^x}{x!}$$

Na przykład prawdopodobieństwo spotkania 4 komórek pasożyta w jednej krwince czerwonej wynosi:

$$P(4) = \frac{e^{-0.193}\,0.193^4}{4!} = \frac{(0.824482)(0.0013874)}{24} = 0.0000476$$

Ponieważ przebadano łącznie 65000 krwinek, to nasza szacowana liczebność oczekiwana wyniesie: 0.0000476 x 65000 = 3.098. W analogiczny sposób możemy obliczyć liczebności oczekiwane dla każdego wariantu liczby komórek zarodźca przypadających na krwinkę.

Liczba estymowanych parametrów rozkładu wynosi 1, zatem *d.f.* = 6 – 1 – 1 = 4. Wartość statystyki $\chi^2_{0.001,4} = 18.47$, czyli możemy z prawdopodobieństwem ponad 99.9% odrzucić hipotezę zerową i uznać, że jest o wiele mniej krwinek z jedną komórką pasożyta, i o wiele więcej krwinek z 3 lub 4 komórkami pasożyta, niż można byłoby oczekiwać przez zwykły przypadek.

Test Kolmogorova-Smirnova

Przykład 115

Wykorzystując test dobroci dopasowania rozkładu Kolmogorova-Smirnova należy określić czy preferencje robotnic osy *Vespula germanica* do odwiedzania żółtych szalek zawierających rozcieńczone roztwory różnych estrów kwasu sorbowego są jednakowe czy różne. Obserwacje uporządkowano w kolejności malejącej polarności podstawnika (alkoholu) w cząsteczce estru.

rodzaj estru	***1***	***2***	***3***	***4***	***5***
liczba robotnic, które przyleciały do szalki w czasie obserwacji	8	13	6	8	5

Wykorzystamy wariant testu Kolmogorova-Smirnova dla danych dyskretnych. Test ten wymaga aby całkowita liczba obserwacji n była równą wielokrotnością liczby kategorii k. Bardzo istotne dla wartości obliczanej statystyki D_{max} jest uporządkowanie danych w kolejności zgodnej z jakąś tendencją. Wartości krytyczne testu podawane są dla przypadków, gdy częstości teoretyczne f_t są równe, ale test sprawdza się także dla nierównych f_t, o ile te różnice nie są duże. Dla niewielkich n test ten ma większą moc niż test chi^2.

	1	2	3	4	5	n
f_i	8	13	6	8	5	40
f_t	8	8	8	8	8	40
$S_N(X)$	8	21	27	35	40	
$F_0(X)$	8	16	24	32	40	
$\lvert D_i \rvert$	0	5	3	3	0	

Maksymalne $\lvert D_i \rvert$ wynosi 5.

$(D_{max})_{0.05.5.40} = 8$

Ponieważ $\lvert D_i \rvert < (D_{max})_{0.05.5.40} = 8$, nie mamy podstaw aby odrzucić hipotezę zerową mówiącą, że rodzaj estru kwasu sorbowego nie ma wpływu na skuteczne przywabianie robotnic *Vespula germanica* do żółtych szalek.

Przykład 116

Korzystając z testu Kolmogorova-Smirnova należy ocenić, czy zagęszczenie osobników włochacza nabrzozka (*Biston betularius* L.) na różnych wysokościach 25-metrowego brzozowego pnia jest równomierne.

X_i	1.4	2.6	3.3	4.2	4.7	5.6	6.4
f_i	1	1	1	1	1	2	1
X_i	7.7	9.3	10.6	11.5	12.4	18.6	22.3
f_i	1	1	1	1	1	1	1

Wykorzystamy wariant testu Kolmogorova-Smirnova dla danych dyskretnych. Przez $F_0(X_i)$ oznaczmy sobie względne skumulowane częstości oczekiwane, a przez $S_N(X_i)$ – względne skumulowane częstości obserwowane. Na określonej wysokości pnia X_i znaleziono f_i osobników motyla. Przez S_{skumul} oznaczmy skumulowane częstości obserwowane – na ich podstawie możemy obliczyć skumulowane względne częstości obserwowane, S_N jako proporcję S_{skumul} na określonej wysokości $\le X_i$: $S_N = \frac{S_{skumul}}{n}$,

gdzie n oznacza liczbę przeprowadzonych pomiarów. Dla każdej wartości X_i 25-metrowego pnia brzozy liczymy skumulowaną częstość oczekiwaną jako:

$$F_{skumul} = \frac{X_i}{25},$$

co oznacza, że badamy rozkład częstości w granicach od 0 do 25 m (a nie od 1 do 25 m). Zatem statystyka dobroci dopasowania będzie wynosić:

$$D_{\max} = \max\left|F_0(X_i) - S_N(X_i)\right|$$

albo $$D'_{\max} = \max\left|F_0(X_i) - S_N(X_{i-1})\right|$$

a nasze $$\max D = \max\left[(D_{\max}), (D'_{\max})\right]$$

co oznacza, że max D jest największą wartością D_{max} lub największą wartością D'_{max}, którakolwiek z nich jest większa.

W naszym przypadku:

i	X_i	f_i	S_{skumul}	S_N	F_0	D_i	D'_i
1	1.4	1	1	0.066667	0.056	0.010667	0.056000
2	2.6	1	2	0.133333	0.104	0.029333	0.037333
3	3.3	1	3	0.200000	0.132	0.068000	0.001333
4	4.2	1	4	0.266667	0.168	0.098667	0.032000
5	4.7	1	5	0.333333	0.188	0.145333	0.078667
6	5.6	2	7	0.466667	0.224	0.242667	0.109333
7	6.4	1	8	0.533333	0.256	0.277333	0.210667
8	7.7	1	9	0.600000	0.308	0.292000	0.225333
9	9.3	1	10	0.666667	0.372	0.294667	0.228000
10	10.6	1	11	0.733333	0.424	0.309333	0.242667
11	11.5	1	12	0.800000	0.460	0.340000	0.273333
12	12.4	1	13	0.866667	0.496	0.370667	0.304000
13	18.6	1	14	0.933333	0.744	0.189333	0.122667
14	22.3	1	15	1.000000	0.892	0.108000	0.041333

$$D_{\max} = \max\left|F_0(X_i) - S_N(X_i)\right| = 0.37067$$

$$D'_{\max} = \max\left|F_0(X_i) - S_N(X_{i-1})\right| = 0.304$$

$$\max D = 0.371$$

Ponieważ $D_{0.05.15} = 0.3376$, hipotezę zerową, mówiącą o równomiernym rozmieszczeniu osobników włochacza nabrzozka (*Biston betularius L.*) na pniu 25-metrowej brzozy, należy odrzucić.

Przykład 117

Czy rodzaj leku przeciwpłytkowego wpływa na czas upływający od przeprowadzenia zabiegu do wystąpienia pierwszego epizodu niedrożności pomostów aortalno-wieńcowych u pacjentów kardiochirurgicznych poddawanych zabiegom z wykorzystaniem krążenia pozaustrojowego? Należy zweryfikować hipotezę, że u pacjentów stosujących lek 2 epizody niedrożności odnotowuje się w takim samym okresie po zabiegu jak u pacjentów zażywających lek 1.

Liczby pacjentów, u których wystąpiły epizody niedrożności*

liczba dni, które upłynęły od zabiegu	lek 1	lek 2
0-2	11	1
3-5	7	3
6-8	8	6
9-11	3	12
12-14	5	12
15-17	5	14
18-20	5	6

* *uwaga: zaleca się aby zastosować tyle kategorii ile możliwe; kryteria klasyfikacji przypadków (zaszeregowywania do kategorii) powinny oczywiście być jednakowe dla obu badanych grup*

Zastosujemy test Kolmogorova-Smirnova do porównania rozkładów dwóch prób badanych. Budujemy hipotezy:
- H_0: czas, który upływa od zabiegu do odnotowania pierwszego epizodu niedrożności nie jest dłuższy u pacjentów stosujących lek 2 w porównaniu z tymi, którzy zażywali lek 1
- H_1: czas, który upływa od zabiegu do odnotowania pierwszego epizodu niedrożności jest dłuższy u pacjentów stosujących lek 2 niż u tych, którzy zażywali lek 1

Procedura obliczeń jest następująca:
1. Obliczyć względne skumulowane częstości obserwowane w obu grupach;
2. Obliczyć różnice między wartościami względnych skumulowanych częstości obserwowanych w obu grupach;
3. Określić maksymalną różnicę, $D_{m,n}$ w przewidywanym kierunku (gdybyśmy stosowali test obustronny, to $D_{m,n}$ oznacza największą różnicę w jakimkolwiek kierunku);
4. Obliczyć wartość statystyki chi^2 wg wzoru:

$$\chi^2 = 4D_{m,n}^2 \frac{mn}{m+n}$$

5. Jeżeli obliczona wartość statystyki D jest większa lub równa wartości tablicowej, to możemy odrzucić hipotezę zerową.
   Obliczenia:

	liczby dni od zabiegu, po których wystąpiły epizody niedrożności						
	0-2	**3-5**	**6-8**	**9-11**	**12-14**	**15-17**	**18-20**
$S_{44}(X)$	$11/44$	$18/44$	$26/44$	$29/44$	$34/44$	$39/44$	$44/44$
$S_{54}(X)$	$1/54$	$4/54$	$10/54$	$22/54$	$34/54$	$38/54$	$54/54$
$S_{44}(X) - S_{54}(X)$	0.232	0.355	0.406	0.252	0.143	0.182	0

Nasze $maxD_{m,n} = 0.406$. Ponieważ zarówno *m*, jak i *n* są większe niż 25, liczymy

$$\chi^2 = 4D_{m,n}^2 \frac{mn}{m+n} = 4(0.406)^2 \frac{(44)(54)}{(44+54)} = (0.6593)\frac{2376}{98} = 15.99$$

Dla $df = 2$ $\chi^2 = 15.99 > \chi^2_{0.001} = 13.816$ dla testu jednostronnego, dlatego też z prawdopodobieństwem ponad 99.9% możemy odrzucić hipotezę zerową i przyjąć hipotezę alternatywną, mówiącą, że czas, który upływa od zabiegu do odnotowania pierwszego epizodu niedrożności jest dłuższy u pacjentów stosujących lek 2 niż u tych którzy zażywali lek 1.

Przykład 118

Dla danych z arkusza ***Kolmogorov.xls*** należy ocenić, czy na podstawie rozkładów danych zmienne LDL-1 i LDL-2 pochodzą z tej samej populacji ogólnej.

Stosujemy test Kolmogorova-Smirnova dla porównania rozkładów dwóch prób.
- H_0: zmienne LDL-1 i LDL-2 pochodzą z tej samej populacji ogólnej, ponieważ nie ma istotnych różnic między rozkładami obu zmiennych,
- H_1: rozkłady obu zmiennych różnią się, czyli zmienne LDL-1 i LDL-2 nie pochodzą z tej samej populacji ogólnej

   Obliczenia:

kategoria	**liczność LDL-1**	$S_{kat}(X_1)$	$S_{201}(X_1)$	**liczność LDL-2**	$S_{kat}(X_2)$	$S_{200}(X_2)$	$S_{201}(X_1)$-$S_{200}(X_2)$
0-20	0	0.000	0.000	0	0.000	0.000	0.000
20-40	0	0.000	0.000	0	0.000	0.000	0.000
40-60	1	0.005	0.005	7	0.035	0.035	0.030
60-80	2	0.010	0.015	14	0.070	0.105	0.090
80-100	16	0.080	0.095	23	0.115	0.220	0.125
100-120	14	0.070	0.164	36	0.180	0.400	0.236
120-140	44	0.219	0.383	26	0.130	0.530	0.147
140-160	36	0.179	0.562	20	0.100	0.630	0.068
160-180	33	0.164	0.726	17	0.085	0.715	0.011

180-200	22	0.109	0.836	11	0.055	0.770	0.066
200-220	16	0.080	0.915	15	0.075	0.845	0.070
220-240	11	0.055	0.970	10	0.050	0.895	0.075
240-260	1	0.005	0.975	4	0.020	0.915	0.060
260-280	3	0.015	0.990	3	0.015	0.930	0.060
280-300	0	0.000	0.990	5	0.025	0.955	0.035
300-320	0	0.000	0.990	4	0.020	0.975	0.015
320-340	0	0.000	0.990	0	0.000	0.975	0.015
340-360	1	0.005	0.995	0	0.000	0.975	0.020
360-380	1	0.005	1.000	0	0.000	0.975	0.025
380-400	0	0.000	1.000	2	0.010	0.985	0.015
400-420	0	0.000	1.000	1	0.005	0.990	0.010
420-440	0	0.000	1.000	1	0.005	0.995	0.005
440-460	0	0.000	1.000	1	0.005	1.000	0.000
Σ	**201**	**1.000**		**200**	**1.000**		**0.236**

Nasze $maxD_{m,n}$ = 0.236. Liczności m = 201 i n = 200 są większe niż 25. zatem obliczamy wartość statystyki χ^2 ze wzoru:

$$\chi^2 = 4D_{m,n}^2 \frac{mn}{m+n} = 4(0.236)^2 \frac{(201)(200)}{(201+200)} = (0.223)\frac{40200}{401} = 22.36$$

Dla $df = 2$ $\chi^2 = 22.36 > \chi^2_{0.001} = 13.816$ dlatego też z prawdopodobieństwem ponad 99.9% możemy odrzucić hipotezę zerową mówiącą, że nie ma istotnych różnic między rozkładami obu zmiennych i że zmienne LDL-1 i LDL-2 pochodzą z tej samej populacji ogólnej: rozkłady zmiennych różnią się.

Rozdział 25

Wybrane metody oparte na teście chi^2 wykorzystywane w badaniach populacyjnych i diagnostycznych

Przykład 119

W badaniach wpływu doustnych środków antykoncepcyjnych (OC) na ryzyko powikłań zakrzepowo-zatorowych wylosowano 215 kobiet z 28 szpitali. Dla każdej z wylosowanych kobiet, u których stwierdzono wstępne oznaki żylnej choroby zakrzepowej, dobrano kobietę cierpiącą na inne schorzenie (o którym wiedziano, że nie jest skojarzone ze stosowaniem OC) z tego samego szpitala i zaklasyfikowano ją jako kontrolę. Zadbano o to, aby pacjentka „kontrolna" miała to samo miejsce zamieszkania (*miasto podobnej wielkości – wieś*), ten sam okres hospitalizacji, rasę, wiek, status małżeński, dochody. Tabela zestawiająca rozkład liczebności jest następująca:

	kontrole		
przypadki	**stosowały OC**	**nie stosowały OC**	***razem***
stosowały OC	73	386	**459**
nie stosowały OC	85	763	**848**
razem	**158**	**1149**	**1307**

$$OR = \frac{386}{85} = 4.54$$

Widzimy, że były 73 pary, w których zarówno kontrole, jak i przypadki stosowały OC (były narażone na ryzyko powikłań zakrzepowych) oraz 763 przypadki, w których zarówno kontrole, jak i przypadki nie stosowały OC (nie były narażone). Te dwa zestawienia liczebności nie niosą nam informacji o zależności między stosowaniem OC a powikłaniami zakrzepowo-zatorowymi. Informacja ta jest rozłożona na pozostałe dwa składniki: 386 par kontroli, które nie stosowały OC z przypadkami, które stosowały OC oraz 85 par przypadków, które nie stosowały OC z kontrolami, które stosowały OC.

OR obliczony dla tych przeciwstawnych par będzie wynosił:

$$\text{OR} = \text{iloraz niezgodnych par} = \frac{liczba\ par: przypadki_{narażone} - kontrole_{nienarażone}}{liczba\ par: przypadki_{nienarażone} - kontrole_{narażone}}$$

OR = 386/85 = 4.54.

Istotność tak obliczonego OR dla par przeciwstawnych zależy od wartości statystyki sparowanego testu McNemara. Wartość ta wynosi:

$$\chi^2 = \frac{(|b-c|-1)^2}{b+c} = \frac{(|386-85|-1)^2}{386+85} = 191.1,$$

i jest ona istotna przy $p< 0.001$.

Przykład 120

Wykorzystując jako miarę zależności współczynnik ryzyka względnego, oszacuj czy palenie papierosów (jako czynnik ryzyka) wpływa na większe ryzyko zawału mięśnia sercowego.

H_0: RR = 1

H_A: RR ≠ 1

	czynnik ryzyka		
wynik	***palący***	***niepalący***	***razem***
zawał	278	93	**371**
bez zawału	128	991	**1119**
razem	**406**	**1084**	**1490**

$$RR = \frac{ab+ad}{ab+ac} = \frac{(278*93)+(278*991)}{(278*93)+(278*128)} = 7.98$$

$$\chi^2 = \frac{n(ad-bc)^2}{(a+b)(b+d)(a+b)(c+d)} = \frac{1490(278*991-93*128)^2}{(406*1084*371*1119)} = 566.63, \qquad p \ll 0.001$$

Zatem

$$\chi = \sqrt{566.63} = 23.8 \qquad -95\%CI = RR^{(1-1.96/\chi)} = 6.73 \qquad +95\%CI = RR^{(1+1.96/\chi)} = 9.47$$

Przedział ufności 95% dla RR wynosi od 6.73 do 9.47, czyli ryzyko zawału u palących jest 6.73 do 9.47 razy większe niż u osób niepalących.

Przykład 121

Gdybyśmy dla danych z poprzedniego przykładu chcieli zbadać wpływ dodatkowego czynnika – płci, to nasza tabela z danymi wyglądałaby tak:

		czynnik ryzyka		
	wynik	***palący***	***niepalący***	***razem***
mężczyźni	***zawał***	135	51	**186**
	bez zawału	61	502	**563**
		196	553	**749**
kobiety	***zawał***	89	120	**209**
	bez zawału	121	411	**532**
	razem	**210**	**531**	**741**

Ponieważ płeć jest dodatkową zmienną, która może wpływać na zależność między paleniem a zawałem, stosujemy model stratyfikacyjny do obliczenia wartości RR:

$$RR_{MH} = \frac{\sum_{i=1}^{k} \frac{a_i(c_i + d_i)}{N_i}}{\sum_{i=1}^{k} \frac{c_i(a_i + b_i)}{N_i}} = \frac{\left[\frac{135(61+502)}{749}\right] + \left[\frac{89(121+411)}{741}\right]}{\left[\frac{61(135+51)}{749}\right] + \left[\frac{121(89+120)}{741}\right]} = 3.356$$

Widzimy, że RR zmalał ponad dwukrotnie, czyli płeć odgrywa istotną rolę w zależności zawału od palenia.

Obliczamy wartość statystyki chi^2 testu Cochrana-Mantela-Haenszela z równania:

$$\chi_{CMH}{}^2 = \frac{\left[\sum (a_i - e_i)\right]^2}{\sum v_i}$$

W tym celu liczymy wartości licznika i mianownika dla każdego z poziomów stratyfikacyjnych:

Dla mężczyzn:

$$e_i = \frac{(a_i + b_i)(a_i + c_i)}{n_i} = \frac{(135+51)(135+61)}{749} = 48.67$$

$$v_i = \frac{(a+b)_i(c+d)_i(a+c)_i(b+d)_i}{(n_i - 1)(n_i{}^2)} =$$

$$= \frac{(135+51)(61+502)(135+61)(51+502)}{(749-1)(749)^2} = 27.05$$

Dla kobiet:

$$e_i = \frac{(a_i + b_i)(a_i + c_i)}{n_i} = \frac{(209)(210)}{741} = 59.23$$

$$v_i = \frac{(a+b)_i(c+d)_i(a+c)_i(b+d)_i}{(n_i - 1)(n_i^2)} = \frac{(209)(532)(210)(531)}{(740)(741)^2} = 30.51$$

	mężczyźni	**kobiety**
a_i	135	89
e_i	48.67	59.23
v_i	27.05	30.51

$$\chi_{CMH}^2 = \frac{\left[\sum (a_i - e_i)\right]^2}{\sum v_i} = \frac{[(135 - 48.67) + (89 - 59.23)]^2}{(27.05 + 30.51)} = \frac{13479.21}{57.56} = 234.18$$

Ponieważ χ_{CMH}^2 jest o wiele większe od $\chi^2_{0.001,1} = 10.83$, to z prawdopodobieństwem ponad 99.9% możemy uznać że dodatnia zależność między paleniem tytoniu a zawałem serca (palenie tytoniu zwiększa ryzyko zawału) zarówno w grupie mężczyzn, jak i kobiet nie jest przypadkowa.

Przykład 122

Badano rozkład polimorfizmu 4G/5G regionu promotorowego genu PAI-1 u pacjentów z chorobą niedokrwienną serca. W grupie ponad 1300 pacjentów z ChNS stwierdzono występowanie allelu 4G (genotypy 4G/4G lub 4G/5G) u ponad 960 pacjentów. Zawał serca wystąpił u ponad 670 osób. Wyniki częstości występowania allelu 4G wśród pacjentów z ChNS, którzy przeżyli zawał, oraz tych, którzy nie mieli zawału, podano poniżej. Należy ustalić, czy występowanie allelu 4G może być czynnikiem ryzyka zawału serca.

Iloraz szans i przedział ufności (*CI*): 4G *vs.* zawał

	4G			
zawał	**0**	**1**		
0	215	431		
1	138	533		
chi^2	*istotność p*	*OR*	*–95% CI*	*+95% CI*
27.122	0.001	1.927	1.505	2.466

Już wstępna ocena częstości wskazuje, że częstość allelu 4G jest większa u pacjentów, u których wystąpił zawał mięśnia sercowego (79% *vs.* 67%). Aby ocenić czy różnica ta jest istotna obliczamy OR, jego istotność oraz przedział ufności:

$$OR = \frac{ad}{bc} = \frac{215 * 533}{431 * 138} = 1.9266$$

oznacza, że prawdopodobieństwo wystąpienia allelu 4G jest prawie dwukrotnie wyższe w grupie pacjentów z zawałem niż u tych, u których zawał nie wystąpił. Oczywiście takie stwierdzenie jest trochę nielogiczne, gdyż zakłada w domyśle, że wystąpienie zawału mogłoby warunkować pojawienie się allelu 4G. Wiemy jednak, że tabela czteropolowa jest symetryczna, dlatego poprawniej będzie sformułować nasz wniosek w ten sposób, że u nosicieli allelu 4G ryzyko wystąpienia zawału będzie prawie dwukrotnie wyższe.

$$\chi^2 = \frac{n(ad-bc)^2}{(a+b)(b+d)(a+b)(c+d)} = \frac{1317(114595-59478)^2}{(646*671*353*964)} = 27.12 \qquad p < 0.001$$

$$\sqrt{\chi^2} = 5.2077$$

OR jest wysoce istotny statystycznie; obliczmy jeszcze przedział ufności dla prawdopodobieństwa 99.9%.
Dla takiego prawdopodobieństwa, wartość $z = 3.29$.
Zatem

$$-99.9\%CI = OR^{(1-3.29/5.208)} = 1.9266^{0.3682431} = 1.27$$

$$+99.9\%CI = OR^{(1+3.29/5.208)} = 1.9266^{1.631757} = 2.915$$

Nasz 99.9% przedział ufności wynosi od 1.27 do 2.915.

Przykład 123

Wśród 500 ochotników, którzy zgłosili się do udziału w badaniach nad skutecznością nowej szczepionki na grypę, wystąpiło 150 przypadków zachorowań na grypę. Rozkład częstości podaje tabela.

grypa	**szczepionka**	**placebo**	***razem***
tak	35	115	**150**
nie	210	140	**350**
razem	**245**	**255**	**500**

Należy oszacować skuteczność szczepionki i przedział ufności dla tego badania.

Już na pierwszy rzut oka widzimy, że znacznie więcej zachorowań wystąpiło wśród osób nieszczepionych (45.1% *vs*. 14.3%). Gdyby osoby niezaszczepione dostały szczepionkę, to zachorowałoby jedynie 14.3% x 255 = 0.143 x 255 = 36.4 osób zamiast 115, czyli uchronilibyśmy przed zachorowaniem 115 – 26.4 = 78.6 osób. Daje to skuteczność szczepionki równą 78.6/115 = 68%.

Ryzyko względne wynosi:

$$RR = \frac{115}{255} : \frac{35}{245} = \frac{(115)(245)}{(255)(35)} = 3.157$$

czyli:

$$\textit{skuteczność szczepionki} = 1 - \frac{1}{3.157} = 1 - 0.31677 = 0.6832 \text{ (jak wyżej)}$$

Wartość statystyki:

$$\chi^2 = 56.511, \text{ zatem } \chi = \sqrt{56.511} = 7.517 \qquad p < 0.001$$

Dolny kres przedziału ufności:

$$-95\%CI = 1 - \frac{1}{RR^{1-1.96/\chi}} = 1 - \frac{1}{(3.157)^{1-1,96/7.517}} = 1 - \frac{1}{2.339} = 0.5725$$

Odpowiednio, górny kres przedziału ufności wynosi:

$$+95\%CI = 1 - \frac{1}{RR^{1+1.96/\chi}} = 1 - \frac{1}{(3.157)^{1+1,96/7.517}} = 1 - \frac{1}{4.26} = 0.7563$$

Zatem 95% przedział ufności dla skuteczności szczepionki wynosi od 57% do 76%.

Miary zapadalności i umieralności

Przykład 124

Rejestrowano zapadalność na odrę w drugim roku życia dzieci w kraju, gdzie nie wykonuje się szczepień przeciwko odrze. Śledzono pojawianie się choroby u 1000 dzieci, począwszy od dnia ich urodzin do ukończenia roku i zaobserwowano, że 140 dzieci zachorowało. Jaka jest zapadalność i częstość choroby w obserwowanej grupie dzieci?

Skoro w pierwszym roku zachorowało 140 dzieci z 1000, to znaczy, że 860 pozostało zdrowych i ciągle narażonych na zachorowanie w ciągu następnych lat.

Ryzyko, że dziecko, które uniknie zachorowania w ciągu pierwszego roku życia i zachoruje w ciągu drugiego (czyli zapadalność w drugim roku) wynosi 140/1000 = 0.14 = 14%. Zapadalność liczona dla pierwszego roku wynosi zresztą tyle samo, gdyż podobnie mamy 1000 dzieci poddanych obserwacji i z tego 140 zapada na odrę. Zapadalność rozumiana jako częstość wystąpienia choroby dotyczy drugiego roku obserwacji. Jeżeli założymy, że zapadalność na odrę jest równomierna w ciągu całego roku, uśredniona liczba dzieci narażonych wynosi (1000+860)/2 = 930. Zatem częstość wystąpienia choroby będzie wynosić 140/930 = 0.1505 = 15.05% czyli 150.5 na 1000 przypadków.

Analiza częstości wystąpienia choroby z wykorzystaniem statystyki rozkładu Poissona

Przykład 125

Badano występowanie infekcji dróg oddechowych u dzieci do lat 5 w zależności od warunków bytowych rodzin. Zaplanowano włączenie do badania 500 dzieci, ale na skutek czynników losowych i pozalosowych (migracje, włączenie nowych urodzeń, ukończenie wieku 5 lat, utrata kontaktu) liczba przebadanych dzieci wzrosła prawie dwukrotnie.

warunki bytowe	liczba infekcji	dzieci-lat narażonych (*pyar*)	częstość wystąpienia choroby/1000 dzieci-lat
złe	46	467	98.5
dobre	32	678	47.2
razem	**78**	**1145**	**68.1**

Jaka jest częstość wystąpienia infekcji w całej grupie dzieci? Należy także policzyć błąd oraz przedział ufności.

Częstość wystąpienia infekcji wyrażona na 1000 dzieci-lat narażonych na infekcję (*pyar*) wynosi:

$$\lambda = 78/1145 \text{ x } 1000 = 68.12 \text{ na } 1000 \text{ dzieci na rok}$$

i odpowiednio:

$$SE = \frac{\sqrt{78}}{1145} x1000 = \sqrt{\frac{68.12}{1145} x1000} = 7.71,$$

zatem 95% przedział ufności ($z = 1.96$) wynosi:

$$68.12 \pm 1.96 \text{ x } 7.71 = 60.41 \text{ do } 75.84 \text{ infekcji na } 1000 \text{ dzieci na rok.}$$

Przykład 126

Na podstawie danych z poprzedniego przykładu należy porównać dwie częstości wystąpienia choroby dla dzieci wychowywanych w złych i dobrych warunkach bytowych.

Już na pierwszy rzut oka widać, że u dzieci żyjących w złych warunkach bytowych częstość wystąpienia choroby jest ponad dwukrotnie wyższa niż u dzieci wychowywanych w dobrych warunkach: odpowiednio 98.5 i 47.2 infekcji na 1000 dzieci na rok.

W celu porównania obu częstości wykorzystamy aproksymację rozkładu Poissona do rozkładu normalnego i obliczymy wartość statystyki z według równania:

$$z = \frac{\left|\lambda_1 - \lambda_2\right| - \left[1/(2pyar_1) + 1/(2pyar_2)\right]}{\sqrt{\left[\lambda(1/pyar_1 + 1/pyar_2)\right]}},$$

We wzorze tym wyrażenie $\lambda = \frac{x_1 + x_2}{pyar_1 + pyar_2}$ = 68.12 jest całkowitą częstością wystąpienia choroby w obu porównywanych grupach łącznie. Częstości są wyrażone na 1000 dzieci na rok, czyli wartości $pyar_1$ i $pyar_2$ będą odpowiednio wynosiły 0.467 i 0.678.

$$z = \frac{|98.5 - 47.2| - [1/(2*0.467) + 1/(2*0.678)]}{\sqrt{[68.12(1/0.467 + 1/0.678)]}} = \frac{|51.3| - [1.808]}{\sqrt{68.12 \times 3.616}} = \frac{49.495}{15.695} = 3.15$$

Różnica między częstościami jest istotna z prawdopodobieństwem większym niż 99.8% (p<0.002), czyli dzieci żyjące w złych warunkach bytowych są istotnie bardziej narażone na infekcje dróg oddechowych.

Procedury standaryzacyjne miar zapadalności i umieralności

Przykład 127

W tabeli na str. 398 przedstawiono chorobowość w dwóch izolowanych podgórskich osadach. Należy policzyć czy chorobowość w obu osadach jest istotnie różna.

Widzimy, że chorobowość wzrasta z wiekiem oraz jest wyższa u osobników płci męskiej. Do standaryzacji ze względu na wiek i płeć wykorzystamy metodę bezpośrednią. Jako populację standardową wybierzemy ogólną populację obejmującą wszystkich mieszkańców osady A i wszystkich mieszkańców osady B.

Obliczamy liczbę osób w populacji standardowej, którzy zachorowaliby gdyby chorobowość z uwzględnieniem wpływu wieku i płci była identyczna jak w wiosce A. Na przykład dla mężczyzn w wieku 30-49 lat 9 z 13 osób zachorowało w wiosce A, co daje wskaźnik chorobowości równy 69.23%. W naszej populacji standardowej jest 13+12 = 25 mężczyzn w takim przedziale wieku, i jeśli chorobowość wynosiłaby także 69.23%, to znaczy, że 17.3 z 25 osób zachorowałoby. W analogiczny sposób możemy obliczyć wartości dla innych grup wiekowych obu płci. W sumie, można spodziewać się, że 140 osób z wioski A oraz 222 osoby z wioski B zachorowałoby, gdybyśmy uwzględnili standaryzację dla wieku i płci, co daje nam chorobowość 140/445x100% = 31.46% dla wioski A i 222/445x100% = 49.89% dla wioski B. Tak skorygowane na wiek i płeć wskaźniki chorobowości dla obu wiosek możemy porównać stosując test chi^2 Mantela-Haenszela. Wartość statystyki χ^2 = 23.78 jest o wiele większa od tablicowej wartości $\chi^2_{0.001,1} = 10.83$, czyli z prawdopodobieństwem ponad 99.9% możemy stwierdzić, że chorobowości wioskach A i B są istotnie różne.

Przykład 128

W tabeli na stronie 399 przedstawiono wyniki umieralności z powodu zawału mięśnia sercowego u nosicieli różnych wariantów polimorfizmu glikoproteiny płytkowej $Pl^{A1/A2}$. Należy ocenić, czy występowanie wariantu $Pl^{A2}(+)$ związane jest z wyższym ryzykiem zgonu z powodu zawału serca.

Oczywiście częstość zawału ma istotny związek z płcią oraz wiekiem badanych: ryzyko wystąpienia zawału u osób starszych, szczególnie płci męskiej, jest znamiennie wyższe. Będziemy zatem oczekiwać, że wśród osób z zawałem będą przeważać osoby w starszym wieku i że będą to w większości mężczyźni, a także – że niezależnie od wieku i płci – będą w tej grupie przeważać nosiciele polimorfizmu $Pl^{A2}(+)$.

W celu obiektywnego porównania między nosicielami $Pl^{A1/A1}$ oraz $Pl^{A2}(+)$ dokonamy standaryzacji metodą pośrednią. Najpierw powinniśmy zdefiniować populację standardową, która powinna być reprezentatywna pod względem struktury wiekowej i proporcji płci oraz powinna być wystarczająco liczna. W naszym przypadku będzie to populacja osób bez wariantu Pl^{A2}. Obliczone współczynniki dla populacji standardowej zastosujemy później do populacji nosicieli $Pl^{A2}(+)$ w celu policzenia, ile epizodów zawału powinniśmy oczekiwać, gdyby ryzyko zawału było takie samo jak wśród nosicieli wariantu $Pl^{A1/A1}$.

Na przykład, można oczekiwać, że proporcja zgonów z powodu zawału serca wśród 180 nosicieli $Pl^{A2}(+)$ płci męskiej w wieku 40-49 lat wynosiłaby 180x(24/1670) = 2.59, gdyby ryzyko zgonu w grupie nosicieli $Pl^{A2}(+)$ było takie samo, jak w grupie nosicieli $Pl^{A1/A1}$. W sumie należałoby oczekiwać 24.03 zgonów na 1000 osób na rok w grupie nosicieli $Pl^{A2}(+)$ różnej płci w różnym wieku. Tymczasem, faktycznie zaobserwowana liczba zgonów wynosiła 75. Proporcja liczby zgonów zaobserwowanych do liczby zgonów oczekiwanych, jeżeli skorygowane pod względem wieku i płci częstości byłyby takie same, jak w populacji standardowej (tzw. umieralność standaryzowana – *standardized mortality ratio*, SMR) wynosi 75/24.03 = 3.12.

Wyraża ona, ile razy bardziej (>1) (lub mniej, <1) prawdopodobne jest, że osoba w populacji badanej (w naszym przypadku osoba z grupy nosicieli $Pl^{A2}(+)$) umrze w porównaniu z osobą w tym samych wieku i tej samej płci w populacji standardowej.

W naszym przypadku, dla populacji nosicieli $Pl^{A1/A1}$ SMR wynosi 1.0, gdyż to właśnie ta grupa była naszą populacją standardową. Dla nosicieli wariantu polimorfizmu $Pl^{A2}(+)$ SMR = 3.12, czyli nosiciel allelu Pl^{A2} jest narażony ponad 3-krotnie bardziej.

Do porównania grup $Pl^{A1/A1}$ i $Pl^{A2}(+)$ wykorzystujemy test chi² Mantela-Haenszela. Wartość $\chi^2 = 57.51 > \chi^2_{0.001,1} = 10.83$, a więc hipotezę o równości współczynników umieralności odrzucamy. Możemy zatem stwierdzić, że niezależnie od płci i wieku nosiciele allelu Pl^{A2} są bardziej narażeni na zgon z powodu zawału mięśnia sercowego.

Współczynniki umieralności skorygowane pod względem proporcji płci i struktury wiekowej można obliczyć na podstawie nieskorygowanych wartości tych miar:

umieralność skorygowana ze względu na wiek i płeć =
= SMR x umieralność nieskorygowana

W naszym przypadku dla populacji nosicieli $Pl^{A1/A1}$ oraz $Pl^{A2}(+)$ będą one wynosić: 13.2 oraz 41.2.

płeć	grupa wiekowa	populacja wioski A			populacja wioski B			populacja standardowa (wioski A+B)		
		liczba przypadków	zachorowało	chorobowość	liczba przypadków	zachorowało	chorobowość	liczba przypadków	skorygowana liczba zachorowań dla wioski A	skorygowana liczba zachorowań dla wioski B
mężczyźni										
	0-4	32	2	6.25	12	2	16.67	44	2.75	7.33
	5-9	44	4	9.09	31	4	12.90	75	6.82	9.68
	10-14	11	3	27.27	9	9	100.0	20	5.45	20.00
	15-29	7	5	71.43	10	9	90.00	17	12.14	15.30
	30-49	13	9	69.23	12	12	100.0	25	17.31	25.00
	>50	22	18	81.82	7	7	100.0	29	23.73	29.00
kobiety										
	0-4	34	1	2.94	19	2	10.53	53	1.56	5.58
	5-9	53	4	7.55	14	3	21.43	67	5.06	14.36
	10-14	16	3	18.75	9	4	44.44	25	4.69	11.11
	15-29	22	14	63.64	11	10	90.91	33	21.00	30.0
	30-49	25	17	68.00	21	20	95.24	46	31.28	43.81
	>50	8	6	75.00	3	3	100.0	11	8.25	11.0
razem		**287**	**86**	**30.00**	**158**	**85**	**53.80**	**445**	**140.0**	**222.20**
skorygowana chorobowość									**31.46%**	**49.89%**

płeć	nosiciele wariantu PI$^{A1/A1}$				nosiciele wariantu PI$^{A2(+)}$				oczekiwana liczba zgonów wśród nosicieli PI$^{A2(+)}$, jeżeli częstości byłyby takie same jak wśród nosicieli PI$^{A1/A1}$
	grupa wiekowa	liczba przypadków	liczba zgonów	zgonów/ 1000/rok	liczba przypadków	liczba zgonów	proporcja nosicieli PI$^{A2(+)}$	zgonów/ 1000/rok	
mężczyźni									
	30-39	2560	21	8.20	130	4	4.8	30.77	1.07
	40-49	1670	24	14.37	180	9	9.7	50.00	2.59
	50-59	1450	20	13.79	255	12	15.0	47.06	3.52
	>60	720	25	34.72	265	26	26.9	98.11	9.20
kobiety									
	30-39	2860	23	8.04	78	3	2.7	38.46	0.63
	40-49	1680	21	12.50	59	2	3.4	33.90	0.74
	50-59	840	18	21.43	174	8	17.2	45.98	3.73
	>60	240	7	29.17	88	11	26.8	125.00	2.57
razem		**12020**	**159**	**13.20**	**1229**	**75**	**9.3**	**61.00**	**24.03**
SMR				1.0 (149/149)				3.12 (75/24.03)	
umieralność skorygowana ze względu na wiek i płeć				13.2				41.2 (3.12 *x* 13.2)	

Czułość, swoistość, predykcja wyników

Przykład 129

Oblicz czułość, swoistość i prawdopodobieństwo wyniku fałszywie ujemnego dla testu ProC®Global wykonywanego u nosicieli mutacji Leiden.

	status genetyczny		
wyniki testu	**mutacja Leiden**	**„typ dziki"**	***razem***
dodatni	74	4	**78**
ujemny	1	71	**72**
razem	**75**	**75**	**150**

$$czułość\ \frac{a}{a+c} = \frac{74}{75} = 0.98676$$

$$swoistość = \frac{d}{b+d} = \frac{71}{75} = 0.9467$$

$$prawdopodobieństwo\ wyniku\ fałszywie\ ujemnego\ = \frac{c}{a+c} = \frac{1}{75} = 0.0133$$

Przykład 130

W tabeli przedstawiono wyniki oporności na aktywowane białko C wykonane przy użyciu metody zmodyfikowanej (z rozcieńczaniem osocza badanego osoczem deficytowym pod względem czynnika V) u nosicieli mutacji Leiden czynnika V. Należy określić podstawowe miary jakościowe wyników oraz wartości predykcyjne, przyjmując, że chorobowość (częstość mutacji Leiden) w naszym regionie geograficznym wynosi 4%.

	wartości rzeczywiste		***razem***
wyniki testu	**mutacja Leiden**	**„typ dziki"**	
oporność na aktywowane białko C	118	14	**132**
brak oporności na aktywowane białko C	2	174	**176**
razem	***120***	***188***	***308***

$$czułość = \frac{a}{a+c} = \frac{118}{120} = 0.983$$

$$swoistość = \frac{d}{b+d} = \frac{174}{188} = 0.926$$

$$\textit{szansa wyniku fałszywie dodatniego} = \frac{b}{b+d} = \frac{14}{188} = 0.074$$

$$\textit{szansa wyniku fałszywie ujemnego} = \frac{c}{a+c} = \frac{2}{120} = 0.017$$

$$PWU = \frac{\left[(1-\textit{czułość})\times \textit{chorobowość}\right]}{\left[(1-\textit{czułość})\times \textit{chorobowość}\right]+\left[\textit{czułość}\times(1-\textit{chorobowość})\right]}$$

$$PWU = \frac{\left[(1-0.983)\times 0.04\right]}{\left[(1-0.983)\times 0.04\right]+\left[0.983\times(1-0.04)\right]} = 0.072\,\%$$

$$PWD = \frac{\left[\textit{czułość}\times \textit{chorobowość}\right]}{\left[\textit{czułość}\times \textit{chorobowość}\right]+\left[(1-\textit{czułość})\times(1-\textit{chorobowość})\right]}$$

$$PWD = \frac{\left[0.983\times 0.04\right]}{\left[0.983\times 0.04\right]+\left[(1-0.983)\times(1-0.04)\right]} = 0.707 \quad \text{czyli } 71\%$$

Widzimy, że test charakteryzuje się bardzo wysoką czułością (w 98 przypadkach na 100 wykrywa mutację Leiden – wykrywa występowanie zmiany genetycznej na podstawie badania funkcjonalnego), dość wysoką swoistością (w 93 przypadkach na 100 jego wynik odpowiada prawdzie), nie dostarcza wielu fałszywie ujemnych wyników (nawet nie 2 na 100), czyli jest niewielkie ryzyko, że przeoczymy występowanie nieprawidłowości, rzadko wskazuje na występowanie cechy, gdy naprawdę jej nie ma (jego wyniki są „niepotrzebnie niepokojące" nawet nie w 8 przypadkach na 100). Zważywszy na częstość występowania mutacji w populacji ogólnej (4%), przewidywanie (predykcja) wyniku dodatniego wynosi około 70%, zaś wyniku ujemnego około 0.1%. Wartości te oznaczają, że prawdopodobieństwo, iż występuje choroba w przypadku, gdy wynik testu jest ujemny (przewidywana wartość ujemna, wartość predykcji wyników ujemnych) wynosi niecałe 0.1%, czyli szansa, że **nie występuje** choroba w przypadku, gdy wynik testu jest ujemny jest bliska 100%. Z kolei, ponieważ przewidywana wartość dodatnia (wartość predykcji wyników dodatnich) wynosi ponad 70%, to prawdopodobieństwo, że **nie występuje** choroba w przypadku, gdy wynik testu jest dodatni wynosi około 30%.

Przykład 131

U 408 osób w starszym wieku, wśród których znajdowały się osoby z cukrzycą typu 2 (n = 84), osoby z nietolerancją glukozy oraz osoby bez potwierdzonych zaburzeń metabolizmu węglowodanów, wykonano oznaczenia stężenia glukozy we krwi na czczo za pomocą przenośnego „glukometru z elektrodami enzymatycznymi" (*blood glucose sensor electrodes*). Na podstawie uzyskanych wyników należy ocenić parametry trafności diagnostycznej rozpoznawania u tych osób cukrzycy, jeżeli przyjmiemy, że wartość progowa dla naszego rozpoznania wynosi 128 mg/dl.

Dla grupy osób bez cukrzycy (n = 324):

96	73	94	**129**	107	104	125	84	82	77
126	**130**	76	78	85	78	108	106	99	76
107	99	116	127	114	85	124	104	51	97
114	64	71	124	83	96	99	60	87	73
94	71	123	**137**	81	108	113	106	80	83
82	97	78	107	**134**	101	104	112	115	96
100	96	119	109	99	117	98	90	110	109
111	90	99	103	107	69	102	110	43	72
81	99	98	70	89	73	61	83	92	86
80	108	113	125	100	101	89	114	104	101
95	**141**	110	84	91	98	**131**	94	70	27
112	100	105	120	72	71	98	110	99	84
81	123	115	107	121	65	107	99	125	86
86	73	117	101	116	**132**	**170**	111	109	115
69	94	119	114	97	110	99	97	127	126
93	73	94	111	119	**153**	125	113	122	70
76	103	78	94	78	92	128	105	89	87
101	124	**138**	115	**136**	92	42	99	109	91
125	121	119	121	82	72	111	97	89	103
131	86	45	**130**	59	109	**133**	106	99	116
51	91	91	81	104	**129**	60	119	103	81
107	95	62	76	103	**134**	89	94	95	**135**
99	86	104	96	109	85	128	104	118	**129**
107	111	99	82	107	120	99	106	110	81
123	77	91	83	124	114	117	93	95	86
80	108	85	117	61	68	95	117	114	99
104	110	93	114	94	95	110	92	82	76
86	95	93	71	80	59	106	87	58	
114	78	108	104	83	106	103	84	74	
89	**146**	114	84	115	112	120	116	125	
82	128	102	**131**	82	112	114	**129**	103	
99	85	94	80	92	91	71	100	98	
118	113	107	118	119	95	83	117	109	

Dla grupy osób z cukrzycą (n = 84):

59	**211**	**191**	54
195	76	**134**	**320**
225	**386**	**179**	**205**
249	**263**	**259**	**229**
122	**142**	49	**248**
310	**242**	**312**	**276**
272	**206**	**131**	**168**
244	**194**	**234**	170
297	**285**	113	**328**
292	**260**	**249**	**221**
257	**306**	**160**	**225**
243	**289**	**275**	**250**
219	**275**	**217**	**252**
172	**173**	97	**373**
304	**210**	**219**	**221**
300	126	**196**	**224**
307	**331**	**239**	94
191	**289**	**276**	**204**
188	**258**	**293**	**223**
270	**150**	128	**316**
289	**235**	**288**	**219**

W grupie osób bez cukrzycy stężenie glukozy we krwi wyższe niż wartość progowa 128 mg/dl miało 21 osób, zaś wśród pacjentów z cukrzycą jedynie u 10 osób stężenie glukozy nie było wyższe niż 128 mg/dl.

Możemy zatem napisać:

	występowanie cukrzycy		
	jest	***brak***	***razem***
wynik testu (+) *(czyli stężenie glukozy we krwi > 128 mg/dl)*	74	21	**95**
wynik testu (-) *(czyli stężenie glukozy we krwi < 128 mg/dl)*	10	303	**313**
razem	**84**	**324**	**408**

Na podstawie tego zestawienia wyliczamy, że:

$$\textit{czułość metody} = \frac{a}{a+c} = \frac{74}{74+10} \text{ (x 100\%) = 88.1\%}$$

$$\textit{swoistość} = \frac{d}{b+d} = 100\% \text{ x (303)/(21+303) = 93.5\%}$$

$$\textit{dokładność rozpoznania} = \frac{a+d}{a+b+c+d} = 100\% \text{ x (74+303)/(74+21+10+303) = 92.4\%}$$

$$\textit{rzeczywista częstość choroby} = \frac{a+c}{a+b+c+d} = 100\% \text{ x (74+10)/(74+21+10+303) = 20.6\%}$$

predykcja dodatnia, czyli przewidywana częstość choroby u osób z wynikiem dodatnim testu

$$= \frac{a}{a+b} = 100\% \text{ x (74)/(74+21) = 77.9\%}$$

predykcja ujemna, czyli przewidywana częstość (stwierdzenia) choroby u osób z wynikiem ujemnym

$$\textit{testu} = \frac{c}{c+d} = 100\% \text{ x (10)/(10+303) = 3.2\%,}$$

i odpowiednio przewidywana częstość wykluczenia choroby u osób z wynikiem ujemnym testu = 100% – 3.2% = 96.8%

$$\textit{szansa wyniku fałszywie dodatniego (wśród osób bez cukrzycy)} = \frac{b}{b+d} = 100\% \text{ x (21)/(324) = 6.5\%}$$

$$\textit{szansa wyniku fałszywie ujemnego (wśród osób z cukrzycą)} = \frac{c}{a+c} = 100\% \text{ x (10)/(84) = 11.9\%}$$

Omówienie

Wyniki obliczeń odzwierciedlają zależności między „jakością" testu diagnostycznego (metody pomiaru) a występowaniem choroby (cukrzycy). Trafność badanej metody, która zależy zarówno od czułości, jak i od swoistości metody, jest wysoka. Metoda pozwala wykryć występowanie choroby u 88 osób na 100 przebadanych (czułość = 88.1%). Z drugiej

strony, 6-7 osób na każde 100 przebadanych zdrowych osób zostanie mylnie sklasyfikowanych jako pacjenci z cukrzycą. W związku z tym, dokładność rozpoznania przy posługiwaniu się tą metodą jest bardzo wysoka: 93.4%. Dalej, prawdopodobieństwo potwierdzenia występowania choroby u osób z dodatnim wynikiem testu (metody) (to znaczy szansa, że osoba zaklasyfikowana jako chora naprawdę choruje na cukrzycę) wynosi niecałe 80%, zaś prawdopodobieństwo wykluczenia choroby na podstawie badania stężenia glukozy jest bardzo wysokie (prawie 97%). Zwróćmy uwagę, że obie te wartości predykcyjne wynikają z charakterystyki samego testu: odpowiadają na pytanie jaka jest szansa, że osoba poddana badaniu przy użyciu danej metody cierpi w rzeczywistości na cukrzycę, biorąc pod uwagę trafność metody badania stężenia glukozy za pomocą testowanego glukometru. Szansa, że uzyskamy wynik powyżej 128 mg/dl wśród osób nie chorujących na cukrzycę wynosi 6.5%, natomiast szansa, że wynik będzie mniejszy niż 128 mg/dl wśród pacjentów z cukrzycą jest prawie dwukrotnie wyższa i wynosi 11.5%. Porównajmy te wartości z obliczoną czułością i swoistością metody. Na podstawie takiego zestawienia moglibyśmy powiedzieć, że czułość metody to inaczej szansa wyniku „nie-ujemnego" (czyli dodatniego) wśród osób chorych na cukrzycę, a swoistość to szansa wyniku „nie-dodatniego" (czyli ujemnego) wśród osób bez cukrzycy. Gdyby częstość występowania cukrzycy w badanej grupie była bardzo niska, np. 0.01% (to znaczy gdybyśmy świadomie dobrali grupę, w której nie byłoby prawie osób chorych na cukrzycę), to także wartość predykcyjna dodatnia byłaby o wiele niższa (około 12.1%). Gdybyśmy natomiast przeprowadzili to badanie w grupie składającej się niemal wyłącznie z pacjentów z cukrzycą (np. 98% badanej grupy stanowiliby chorzy), to przewidywana częstość wykrycia choroby u osoby z wynikiem poniżej 128 mg/dl wynosiłaby 99.8% (czyli prawdopodobieństwo wykluczenia choroby u tych osób wynosiłoby około 0.2%), zatem niemal wszystkie (98%) wyniki ujemne (poniżej 128 mg/dl) byłyby fałszywie ujemne.

Przykład 132

Spośród danych analizowanych w poprzednim przykładzie wybierzmy jedynie te przypadki, które mieszczą się w zakresie 60-250 mg/dl glukozy we krwi pełnej. Uzyskujemy tabelę liczebności:

	występowanie cukrzycy		
	jest	***brak***	***razem***
wynik testu (+) *(czyli stężenie glukozy we krwi > 128 mg/dl)*	41	20	**61**
wynik testu (-) *(czyli stężenie glukozy we krwi < 128 mg/dl)*	8	297	**305**
razem	**49**	**317**	**366**

Obliczone wskaźniki trafności diagnostycznej przyjmą wartości:

$$czułość\ metody = \frac{a}{a+c} = 100\% \text{ x } (41)/(41+8) = 83.7\%$$

$$swoistość = \frac{d}{b+d} = 100\% \text{ x } (297)/(20+297) = 93.7\%$$

$$\textit{dokładność rozpoznania} = \frac{a+d}{a+b+c+d} = 100\% \text{ x } (41+297)/(366) = 92.3\%$$

$$\textit{rzeczywistą częstość choroby} = \frac{a+c}{a+b+c+d} = 100\% \text{ x } (41+8)/(366) = 13.4\%$$

$$\textit{predykcja dodatnia} = \frac{a}{a+b} = 100\% \text{ x } (41)/(41+20) = 67.2\%$$

$$\textit{predykcja ujemna} = \frac{c}{c+d} = 100\% \text{ x } (8)/(8+297) = 2.6\%,$$

i odpowiednio, przewidywana częstość wykluczenia choroby u osób z wynikiem ujemnym testu = 100% - 2.6% = 97.4%

$$\textit{szansa wyniku fałszywie dodatniego} = \frac{b}{b+d} = 100\% \text{ x } (20)/(317) = 6.3\%$$

$$\textit{szansa wyniku fałszywie ujemnego} = \frac{c}{a+c} = 100\% \text{ x } (8)/(49) = 16.3\%$$

Jeżeli dane te przedstawilibyśmy na skali przedziałowej, to uzyskamy następujące częstości w każdej z grup dla odpowiednich wartości progowych dla osób zdrowych i chorych na cukrzycę:

		zdrowi		
górny kres przedziału	**liczność**	**częstość skumulowana [%]**	**< górnego kresu przedziału**	**> górnego kresu przedziału**
60	2	0.63%	2	315
70	12	4.42%	14	303
80	31	14.20%	45	272
90	47	29.02%	92	225
100	67	50.16%	159	158
110	59	68.77%	218	99
120	53	85.49%	271	46
130	31	95.27%	302	15
140	11	98.74%	313	4
150	2	99.37%	315	2
160	1	99.68%	316	1
170	1	100.00%	317	0
180	0	100.00%	317	0
190	0	100.00%	317	0
200	0	100.00%	317	0
210	0	100.00%	317	0
220	0	100.00%	317	0
230	0	100.00%	317	0
240	0	100.00%	317	0
250	0	100.00%	317	0

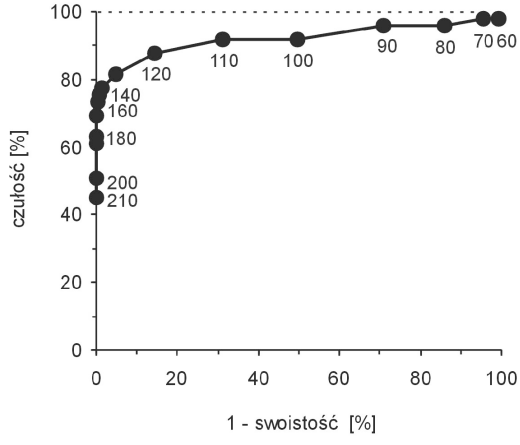
cukrzyca

górny kres przedziału	**liczność**	**częstość skumulowana [%]**	**< górnego kresu przedziału**	**> górnego kresu przedziału**
60	1	2.04%	1	48
70	0	2.04%	1	48
80	1	4.08%	2	47
90	0	4.08%	2	47
100	2	8.16%	4	45
110	0	8.16%	4	45
120	2	12.24%	6	43
130	3	18.37%	9	40
140	2	22.45%	11	38
150	1	24.49%	12	37
160	1	26.53%	13	36
170	2	30.61%	15	34
180	3	36.73%	18	31
190	1	38.78%	19	30
200	5	48.98%	24	25
210	3	55.10%	27	22
220	6	67.35%	33	16
230	7	81.63%	40	9
240	3	87.76%	43	6
250	6	100.00%	49	0

Rozważmy, sytuację, w której każdą granicę przedziału uważalibyśmy za potencjalną wartość progową. Przy takim założeniu dla każdej wartości progowej policzmy czułość oraz swoistość stosowanej metody pomiaru wśród pacjentów z cukrzycą oraz osób nie chorujących na cukrzycę.

górny kres przedziału	***czułość [%]***	***swoistość [%]***	***1-swoistość [%]***
60	98.0	0.6	99.4
70	98.0	4.4	95.6
80	95.9	14.2	85.8
90	95.9	29.0	71.0
100	91.8	50.2	49.8
110	91.8	68.8	31.2
120	87.8	85.5	14.5
130	81.6	95.3	4.7
140	77.6	98.7	1.3
150	75.5	99.4	0.6
160	73.5	99.7	0.3
170	69.4	100.0	0.0
180	63.3	100.0	0.0
190	61.2	100.0	0.0
200	51.0	100.0	0.0
210	44.9	100.0	0.0
220	32.7	100.0	0.0
230	18.4	100.0	0.0
240	12.2	100.0	0.0
250	2.0	100.0	0.0

Zauważmy, że na krańcach rozkładu dobranych przez nas wartości progowych wysoka czułość odpowiada niskiej swoistości i odwrotnie. Zależność między czułością a swoistością metody możemy przedstawić graficznie jako krzywą typu ROC, opisującą zmiany czułości testu w zależności od wyrażenia (1-swoistość). Dla naszego przykładu krzywa ta ma postać:



O dokładności badanej metody możemy wnioskować na podstawie wygięcia krzywej ROC w górnej lewej części wykresu oraz wielkości pola pod krzywą: im krzywa ta jest bardziej wygięta w kierunku górnego lewego rogu, czyli im większa powierzchnia wyznaczana przez 100% czułości i 100% wartości zmiennej (1-swoistość) przypada poniżej krzywej, tym metoda jest bardziej dokładna (bardziej czuła i bardziej swoista). I odwrotnie, im bardziej krzywa zbliżona jest do przekątnej układu współrzędnych, tym test jest mniej dokładny. W naszym przykładzie pole powierzchni pod krzywą stanowi około 85% całości wyznaczanej przez 100% czułości i 100% wartości zmiennej x (1-swoistość), zatem możemy wnioskować, że dokładność metody oznaczania stężenia glukozy we krwi pełnej przy użyciu „glukometru z elektrodami enzymatycznymi", wyznaczona graficzną metodą krzywej ROC wynosi około 85%.

Przykład 133

Stopień asymetrii fosfolipidów błon płytek krwi uznaje się za wskaźnik właściwości prokoagulacyjnych płytek: w płytkach spoczynkowych asymetria ta jest wysoka, zaś po aktywacji komórek postępuje symetryzacja rozmieszczenia cząsteczek fosfolipidów, co wiąże się z nabywaniem przez płytki własności prokoagulacyjnych. Do badania asymetrii rozmieszczenia fosfolipidów błon płytek krwi aktywowanych różnymi agonistami zastosowano dwie metody: znakowania merocjaniną 540 (MC540) oraz wiązania aneksyny V.

	asymetria niezmieniona		asymetria obniżona	
	aneksyna V	***MC540***	***aneksyna V***	***MC540***
średnia	**10.3**	90.1	44.1	59.3
	3.0	77.1	19.0	26.7
	1.9	84.3	23.6	39.4
	0.8	86.7	43.7	37.7
	1.5	85.1	11.5	54.6
	1.6	88.9	13.7	**65.6**
	1.2	79.6	14.3	40.6
	3.2	92.1	15.4	77.0

4.6	**57.9**	36.7	29.8
8.7	92.2	16.2	**66.9**
1.6	**36.9**	58.5	20.9
3.8	84.4	57.3	49.7
0.3	**55.4**	24.5	43.6
3.0	66.6	19.6	34.6
3.8	79.8	30.8	33.5
3.8	80.9	20.8	57.9
1.5	87.5	14.9	59.7
1.1	89.9	29.9	32.1
1.4	94.0	**9.0**	**76.3**
1.8	**55.2**	45.3	15.4
2.2	87.6	31.3	45.9
2.4	93.2	40.4	33.5
5.6	78.2	45.3	**70.6**
7.1	**33.4**	24	50.5
5.3	91.8	11.3	63.4
4.9	78.7	18.8	49.3
7.0	84.8	26.2	**73.8**
8.3	84	12.6	55.7
4.7	62.9	47.3	37.0
5.0	89.1	22.1	**64.7**
5.0	**42.6**	31.1	35.2
		42.2	**72.5**
		53.4	28.5
		40.4	41.9
		20.1	55
		35.4	44.8
		34.8	28.1
		40.2	27.0
		26.3	45.5
		45.1	6.1
		29.0	**66.5**
		51.2	19.4
		37.4	47.3
		53.8	19.1
		46.7	29.7
		46.8	25.5
		24.0	32.6
		25.3	49.9
		50.2	27.6
		41.6	57.1
		42.5	17.3
		42.7	45.4
		76.1	16.7
średnia	*77.1*	*33.3*	*43.5*
SD	*17.1*	*14.9*	*17.8*

* *wartości nietypowe wyróżniono*

Dla potrzeb niniejszego opracowania przyjmijmy, że wartość progowa wynosi dla aneksyny V: dla płytek z normalną asymetrią fosfolipidów (brak aktywności prokoagulacyjnej) – poniżej 10%, dla płytek z asymetrią obniżoną (wykazujących własności prokoagulacyjne) – powyżej 10%, oraz dla MC540: dla płytek z normalną asymetrią fosfolipidów (brak aktywności prokoagulacyjnej) – powyżej 60%, dla płytek z asymetria obniżoną (wykazujących własności prokoagulacyjne) – poniżej 60%.

Korzystając z metody krzywych ROC należy stwierdzić, która z dwóch stosowanych metod oznaczania asymetrii fosfolipidów błon aktywowanych płytek krwi jest bardziej dokładna.

Zestawiamy tabele częstości dla każdej z badanych metod. Dla metody znakowania aneksyną tabela częstości ma postać:

	asymetria fosfolipidów		
	zmieniona	***niezmieniona***	***razem***
wynik testu (+) *(czyli wiązanie aneksyny V > 10%)*	52	1	**53**
wynik testu (-) *(czyli wiązanie aneksyny V < 10%)*	1	30	**31**
razem	**53**	**31**	**84**

Obliczone wskaźniki czułości i swoistości wynoszą:

$$czułość\ metody = \frac{a}{a+c} = 100\% \text{ x } (52)/(52+1) = 98.1\%$$

$$swoistość = \frac{d}{b+d} = 100\% \text{ x } (30)/(30+1) = 96.8\%$$

Dla metody z MC540 tabela taka wygląda następująco:

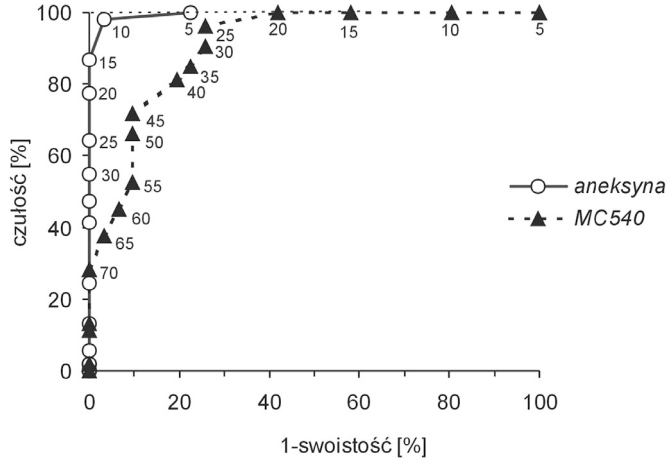
	asymetria fosfolipidów		
	zmieniona	***niezmieniona***	***razem***
wynik testu (+) *(czyli wiązanie MC540 < 60%)*	45	6	**51**
wynik testu (-) *(czyli wiązanie MC540 > 60%)*	8	25	**33**
razem	**53**	**31**	**84**

a obliczona czułość i swoistość metody wynoszą odpowiednio 84.9% i 80.6%.

Zestawiamy tabelę czułości i swoistości każdej z metod w zależności od dobranych wartości progowych. W przypadku metody z MC540 stopień symetryzacji fosfolipidów błony szacujemy jako (100 - wiązanie MC540)%, czyli wartość progowa stopnia symetryzacji będzie wynosić 40%. Odpowiednio, dla płytek z normalną asymetrią fosfolipidów (brak aktywności prokoagulacyjnej) – oczekujemy wartości stopnia symetryzacji poniżej 40%, zaś dla płytek z asymetrią obniżoną (wykazujących własności prokoagulacyjne) – oczekujemy wartości stopnia symetryzacji powyżej 40%.

	aneksyna		MC540	
górny kres przedziału	**1-swoistość [%]**	**czułość [%]**	**1-swoistość [%]**	**czułość [%]**
5	22.6	100.0	100.0	100.0
10	3.2	98.1	80.6	100.0
15	0.0	86.8	58.1	100.0
20	0.0	77.4	41.9	100.0
25	0.0	64.2	25.8	96.2
30	0.0	54.7	25.8	90.6
35	0.0	47.2	22.6	84.9
40	0.0	41.5	19.4	81.1
45	0.0	24.5	9.7	69.8
50	0.0	13.2	9.7	66.0
55	0.0	5.7	9.7	52.8
60	0.0	1.9	6.5	45.3
65	0.0	1.9	3.2	37.7
70	0.0	1.9	0.0	28.3
75	0.0	1.9	0.0	13.2
80	0.0	0.0	0.0	11.3
85	0.0	0.0	0.0	1.9
90	0.0	0.0	0.0	1.9
95	0.0	0.0	0.0	0.0
100	0.0	0.0	0.0	0.0

Na podstawie wykresu przedstawiającego krzywe ROC dla obu zastosowanych metod oceny stopnia symetryzacji fosfolipidów dwuwarstwy lipidowej błon aktywowanych płytek wnioskujemy, że o wiele bardziej dokładnym testem jest metoda oznaczania stopnia wiązania przez płytki aneksyny V.



Niemal cała powierzchnia wyznaczana przez 100% czułości i 100% wartości zmiennej *x* (1-swoistość) mieści się pod krzywą ROC dla aneksyny. Z drugiej strony, kształt krzywej ROC charakteryzującej metodę znakowania płytek MC540 wyraźnie „odstaje" w lewej górnej części wykresu, zatem na podstawie analizy graficznej z wykorzystaniem metody krzywych ROC wnioskujemy, że test ten odznacza się o wiele niższą czułością i swoistością.

Literatura

Podręczniki o zastosowaniu ogólnym:

Afifi A.A., Clark V.: Computer-aided multivariate analysis, 2nd ed., Van Nostrand Reinhold, New York-London-Melbourne, 1990, 463 s.

Armitage P.: Metody statystyczne w badaniach medycznych, PZWL, Warszawa, 1978, 412 s.

Armitage P., Berry G.: Statistical Methods in Medical Research, 3rd ed., Blackwell Science, Oxford-London-Edinburgh, 1994, 630 s.

Bland M.: An Introduction to Medical Statistics, 3rd ed., Oxford Medical Publications, Oxford, 2000, 405 s.

Cochrane W.G., Cox G.M.: Experimental Designs, 2nd ed., John Wiley & Sons, Inc., London, 1992, 617 s.

De Muth J.E.: Basic Statistics and Pharmaceutical Statistical Applications. Marcel Dekker, Inc., New York-Basel, 1999, 596 s.

Field A.: Discovering Statistics Using SPSS (and sex, drugs and rock'n'roll), 2nd ed., Sage Publications, London-Thousand Oaks-New Delhi, 2005, 780 s.

Fisher R.A.: Statistical Methods for Research Workers, Oliver and Boyd, Edinburgh: Tweeddle Court-London, 1925, 239 s.

Fisher R.A.: The Logic of Inductive Inference. J. Royal Stat. Soc. 1935, 98, 39-82.

Francuz P., Mackiewicz R.: Liczby nie wiedzą, skąd pochodzą – Przewodnik po metodologii i statystyce nie tylko dla psychologów. KUL, Lublin, 2007, 654 s.

Good P.I., Hardin J.W.: Common Errors in Statistocs (and How to Avoid them). 2nd ed., John Wiley & Sons, Inc., Hoboken, NJ, 2006, 254 s.

Hill T., Lewicki P.: Statistics – Methods and Applications: A Comprehensive Reference for Science, Industry, and Data Mining. Statsoft, Tulsa, OK, 2006, 832 s.

Katz M.H.: Multivariable Analysis – A Practical Guide for Clinicians. Cambridge University Press, Cambridge, 2001, 192 s.

Lang T.A., Secic M.: How to report statistics in medicine. Annotated Guidelines for Authors, Editors, and Reviewers. ACP, Philadelphia, 1997, 367 s.

Siegel S. and Castellan N.J. Jr.: Nonparametric statistics for the behavioral sciences. 2nd ed., New York, McGraw-Hill Book Company, 1988, 399 s.

Sokal R.R., Rohlf F.J.: Biometry – The principles and practice of statistics in biological research. 2nd ed., San Francisco, W.H. Freeman & Co., 1981, 862 s.

Stanisz A.: Biostatystyka. Wyd. Uniwersytetu Jagiellońskiego, Kraków, 2005, 410 s.

Stanisz A.: Przystępny kurs statystyki z zastosowaniem STATISTICA PL na przykładach z medycyny, tom 1. Statystyki podstawowe. StatSoft Polska, Kraków, 2006, 532 s.

Stanisz A.: Przystępny kurs statystyki z zastosowaniem STATISTICA PL na przykładach z medycyny, tom 2. Modele liniowe i nieliniowe, StatSoft Polska, Kraków, 2007, 868 s.

Stanisz A.: Przystępny kurs statystyki z zastosowaniem STATISTICA PL na przykładach z medycyny, tom 3. Analizy wielowymiarowe. StatSoft Polska, Kraków, 2007, 500 s.

Zar J.: Biostatistical analysis, 4th ed., Prentice-Hall International, Inc. Simon & Schuster/A Viacom Company. Upper Saddle River, N.J., 1999, 663 s.

Zieliński T.: Jak pokochać statystykę czyli STATISTICA do poduszki, StatSoft Polska, Kraków, 1999, 256 s.

Zieliński W.: Tablice statystyczne, seria: Wykłady ze Statystyki i Doświadczalnictwa, wyd. 3, Fundacja „Rozwój SGGW", Warszawa, 1999, 84 s.

Zastosowania szczegółowe oraz wybrane publikacje źródłowe:

Barnett V., Toby L.: Outliers in statistical data., 3rd ed., John Wiley & Sons, Inc., 1994, 1-583.

Bland M.J., Altman D.G.: Measurement Error and correlation coefficients. British Medical Journal 1996, 313, 41-2.

Bland M.J., Altman D.G.: Statistics Notes: Measurement error. British Medical Journal 1996, 312, 1654.

Bland M.J., Altman D.G.: Statistics Notes: Transforming data. British Medical Journal 1996, 312, 770.

Goddard M.J., Hinberg I.: Receiver operating characteristic (ROC) curves and non-normal data: an empirical study. Stat. Med. 1990, 9, 325-337.

Hanley J.A.: Receiver operating characteristic methodology: the state of the art. CRC Critical Reviews in Diagnostic Imaging 1989, 29, 307-335.

Lachenbruch P.A.: Discriminant Analysis. Macmillan Pub. Co., NY: Hafner, 1975, 1-128.

Thompson M.L., Zucchini W.: On the statistical analysis of ROC curves. Stat. Med. 1989, 8, 1277-1290.

Zou K.H., Hall W.J., Shapiro D.: Smooth non-parametric ROC curves for continuous diagnostic tests. Stat. Med. 1997, 16, 2143-56.

Zweig M.H., Campbell G.: ROC plots: a fundamental evaluation tool in clinical medicine Clin. Chem. 1993, 39, 561-577.

Aspekty historyczne statystyki:

Francuz P., Mackiewicz R.: Liczby nie wiedzą, skąd pochodzą – Przewodnik po metodologii i statystyce nie tylko dla psychologów. KUL, Lublin, 2007, 1-654.

Skrabanek P., McCormick J.: Follies & Fallacies in Medicine, 3rd ed., Tarragon Press, Whithorn, 1998, 1-171.

Salsburg D.: The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century. Owl Books, New York, 2002, 1-340.

Bardzo dobre opracowania merytoryczne oraz przykłady zastosowań w zakresie tak podstawowych, jak i zaawansowanych metod statystycznych, znajdzie Czytelnik na platformie firmy StatSoft – dystrybutora programu STATISTICA PL. Godne polecenia są także kursy szkoleniowe – z niezwykle zróżnicowaną tematyką szkoleń – organizowane przez StatSoft Polska.

Indeks