Ryszard Tadeusiewicz
Andrzej Izworski
Janusz Majewski

BIOMETRIA

WYDAWNICTWA AGH —————— KRAKÓW 1993

Wydano za zgodą
Rektora Akademii Górniczo-Hutniczej
im. Stanisława Staszica w Krakowie
w serii skryptów uczelnianych SU 1349

Redaktor Naczelny Uczelnianych Wydawnictw
Naukowo-Dydaktycznych: *prof. dr hab. inż. Zdzisław Kłeczek*

Z-ca Redaktora Naczelnego: *dr inż. Danuta Flisiak*

Recenzent: *prof. dr hab. Tadeusz Grabiński*
Opracowanie edytorskie: *mgr inż. Bożena Dębska*

Skład komputerowy: WYDAWNICTWO FORTIS ul. Urzędnicza 20/6
tel. (012) 34-09-40, 66-87-50

Redakcja Uczelnianych Wydawnictw
Naukowo-Dydaktycznych
al. Mickiewicza 30, paw. A-1, pok. 129, 30-059 Kraków
tel. 33-76-00, 33-81-00, 33-91-00, w. 32-28

ISSN 0239-6114

Spis treści

1. WSTĘP

Termin *biometria* zwiódł już i zmylił wielu badaczy i amatorów. Zanim więc zostaną przedstawione zagadnienia szczegółowe — kilka wyjaśnień zupełnie podstawowej, elementarnej wręcz natury. Otóż wbrew niektórym wyobrażeniom i przypuszczeniom biometria *nie jest* działem metrologii, zajmującym się pomiarami parametrów i cech rozmaitych systemów biologicznych. Takie pomiary i obserwacje są wprawdzie dokonywane na gruncie **anatomii, histologii, antropologii, fizjologii, biofizyki** i w niezliczonej mnogości dalszych szczegółowych działów medycznych badań podstawowych i obserwacji klinicznych, jednak sam proces i metodyka odpowiednich pomiarów nie odbiega w niczym od klasycznych pomiarów wykorzystywanych w technice, fizyce czy chemii.

Rejestrując potencjały elektryczne pracującego serca, siły rozwijane przez mięsień, promieniowanie izotopów metabolizowanych przez wątrobę czy wymiary Miss Polonia — posługujemy się z reguły takimi samymi przyrządami, jak przy pomiarach innych napięć, naprężeń, radiacji czy długości. Oczywiście konieczne jest zbudowanie każdorazowo stosownego stanowiska pomiarowego (na przykład łączącego badany mięsień z dynamometrem) oraz staranny dobór parametrów aparatury zgodnie z ograniczeniami narzuconymi przez naturę badanego zjawiska biologicznego (na przykład konieczność stosowania wzmacniaczy o bardzo dużej impedancji wejściowej przy rejestracji biopotencjałów). Jednak podobne wymagania i ograniczenia występują przy wszelkich pomiarach i obserwacjach przyrodniczych, zatem nawet zebranie i usystematyzowanie wszelkich tego typu zaleceń i przesłanek nie upoważnia jeszcze do stosowania nazwy *biometria* — gdyż będzie to w istocie metrologia.

Tym, co wyróżnia pomiary biologiczne jest natomiast stosunek do wyniku pomiaru. Przy dowolnym pomiarze technicznym lub przy dowolnym doświadczeniu fizycznym badacz ma świadomość, że wyznaczany parametr lub wymiar *obiektywnie istnieje*, a ewentualne niedokładności jego określenia wynikają z niedoskonałości aparatury i zastosowanej metodyki eksperymentu. Jeśli więc stosujemy statystyczną (lub dowolną inną) analizę wyników pomiarów, to głównie traktujemy ją jako analizę błędów i wykorzystujemy do eliminacji niedoskonałości przyrządów pomiarowych. Wielokrotne powtarzanie pomiarów — jeśli jest już w ogóle stosowane — służy polepszeniu jakości pomiaru i może być (przynajmniej w teorii) zastąpione jednorazowym zastosowaniem doskonalszego przyrządu pomiarowego (na przykład miernika wyższej klasy).

Tymczasem w biologii i medycynie już samo mierzone zjawisko jest niepewne. Dokładniej: na każdy pojedynczy pomiar lub na każdą oddzielną obserwację składają się zarówno czynniki, które chcemy kontrolować i analizować, jak i dziesiątki czynników ubocznych, mających niejednokrotnie przemożny wpływ na wynik obserwacji czy pomiaru. Wynika to z samej istoty pomiaru biologicznego. Każdy pacjent, każda żywa istota czy każda rozpatrywana komórka stanowi jedyną i niepowtarzalną indywidualność, posiadającą własne cechy, uwarunkowania i właściwości. Jeśli dokonujemy obserwacji, to spostrzegamy efekt, będący *wypadkową* cechy, której poszukujemy i właściwości indywiduum, które jest tej cechy nosicielem. Jeśli dokonujemy pomiaru, to mierzymy wielkość będącą wynikiem swoistego „przefiltrowania" poszukiwanego parametru przez jednostkowe cechy tego konkretnego osobnika, u którego pomiar jest wykonywany. Jeśli prowadzimy jakikolwiek eksperyment, to mamy w rękach jedynie drobną część czynników wpływających na końcowy efekt, zaś znacznie więcej kart ma w ręce Natura ...

Dlatego *żadna* pojedyncza obserwacja medyczna nie jest miarodajna i *żaden* pomiar biologiczny nie może być traktowany jako dokładny niezależnie od klasy przyrządu, jakim dokonano pomiaru i rzetelności obserwatora, który relacjonował spostrzeżenie. Z tego względu wszystkie eksperymenty i wszystkie obserwacje medyczne, wszelkie pomiary i wszelkie porównania biologiczne muszą się odnosić do zbiorowości. Obserwacje trzeba powtarzać, doświadczenia dublować i rozbudowywać o elementy kontrolne a pomiary wykonywać wielokrotnie u wielu osobników lub w wielu niezależnych próbach.

Jednak tak powielane obserwacje i pomiary — w dodatku z zasady obarczone czynnikami obniżającymi ich wiarogodność — stanowią bardzo niepewną i niewygodną podstawę przy próbach wnioskowania na ich podstawie o właściwościach badanych obiektów i zjawisk a także przy próbach uogólnień i praktycznych zastosowań wyników badań naukowych i obserwacji klinicznych. Dlatego także *niezbędnym* elementem każdego pomiaru i każdej oceny odniesionej do systemów biologicznych musi być *statystyczne opracowanie wyników*. Dzięki takiemu opracowaniu możliwe staje się sprowadzenie wielu mało czytelnych pomiarów do kilku łatwych w interpretacji wskaźników. W dodatku rozsądnie stosowana statystyka daje możliwość precyzyjnego wnioskowania w oparciu o niepewne i obarczone błędami dane. W ten sposób **biometria** jest elementem wydobywającym porządek z chaosu, czynnikiem pozwalającym przezwyciężyć podstawową sprzeczność, jaka istnieje pomiędzy naturą zindywidualizowanych osobniczo obserwacji biologicznych i wywodzącymi się z ideałów nauk ścisłych tendencji do formułowania sądów ogólnych i uniwersalnych.

Reguły wnioskowania statystycznego używane w biometrii, nie są szczególnie wyrafinowane. Używa się metod sprawdzonych i pewnych, wychodząc z założenia, że w sytuacji kiedy może chodzić o zdrowie lub życie ludzkie — to co pewne i niezawodne powinno być wyżej cenione niż to, co błyskotliwie erudycyjne, metodologicznie nowatorskie lub zwyczajnie sprytne. Statystyka jest w biometrii *narzędziem*, co oznacza, że nie będziemy

wnikali w matematyczne podstawy stosowanych metod i pominiemy wszelkie dowody i oparte na aksjomatycznym rachunku prawdopodobieństwa ogólne rozważania, dodające naukowego splendoru większości książek z tej dziedziny, lecz bardzo mało przydatne w praktyce. Czyniąc tak rozczarujemy być może niektórych czytelników, poszukujących w każdej dziedzinie wiedzy ujęć obfitujących w skomplikowane wzory i niezrozumiałe terminy. Być może zestawione na końcu skryptu książki innych autorów, z reguły pisane w sposób bardziej zmatematyzowany, zaspokoją te tęsknoty. W samym skrypcie pokazywać będziemy jedynie, jakie metody są używane w biometrii, do czego służą i jak z nich należy korzystać. Ani mniej — ani więcej. Poznamy narzędzia od strony ich użytkowania, nie zaś od strony ich struktury. Wychodzimy bowiem z założenia, że wprawdzie można i niekiedy trzeba dokonywać analiz krystalograficznych metalu, z którego wykonano młotek, jednak praktykowi najczęściej potrzebna jest jedynie informacja, jak ten młotek uchwycić, w co nim uderzyć oraz — co najważniejsze i najczęściej pomijane — po co uderzyć. Właśnie takiej, na wskroś praktycznej wiedzy dostarcza ten skrypt. Ma on pomagać studentom Elektroniki, specjalizującym się w aparaturze biomedycznej, w poznaniu (i stosowaniu!) zasad biometrii. Specjaliści z tego zakresu są wciąż potrzebni i usilnie poszukiwani we wszystkich większych placówkach Służby Zdrowia, a absolwenci Elektroniki jako dobrze wykształceni matematycznie i dysponujący znaczną biegłością w zakresie używania metod i środków informatyki — mają pełne szanse uzupełnić swoje wykształcenie o te poszukiwane i cenione umiejętności. Znając zasady biometrii mogą oni poprawnie formułować zadania oceny budowanej (lub tylko eksploatowanej...) przez siebie nowoczesnej, elektronicznej aparatury diagnostycznej, terapeutycznej lub służącej potrzebom protetyki i rehabilitacji. Dzięki znajomości biometrii możliwe jest także wspomaganie konsultacją i pomocą obliczeniową badań prowadzonych przez lekarzy i naukowców zatrudnionych w placówkach służby zdrowia, co stale jest w cenie i wciąż występujące tu potrzeby pozostają w tyle za możliwościami ich zaspokojenia.

Jest jednak jeszcze jeden aspekt tej sprawy, godny poruszenia i rozważenia. Pełen rezerwy (oględnie mówiąc) stosunek większości lekarzy i biologów do matematyki jest wręcz przysłowiowy. Na tle tej generalnej niechęci absolutnym wyjątkiem jest bezwzględne zaufanie i powszechna akceptacja (niekiedy przesadna) statystyki. Pojawia się — z gruntu błędna — tendencja do „faszerowania" statystyką wszelkich artykułów i dysertacji biologicznych i medycznych, przy czym wcale częsta jest sytuacja, kiedy obliczenia statystyczne są „sztuką dla sztuki", a nie techniką wnioskowania. Będzie wielką satysfakcją dla autorów skryptu, jeśli jego czytelnicy potrafiąc zasugerować *sensowne* użycie opisywanych w nim metod statystycznych, znajdą w sobie dość zdrowego rozsądku, krytycyzmu i odwagi, żeby przeciwstawiać się „dekoracyjnemu" traktowaniu statystyki w medycynie.

Omówionym wyżej założeniom odpowiada struktura skryptu. W kolejnym, drugim rozdziale zawarte są podstawowe informacje na temat **przygotowania danych** pochodzących z doświadczenia biologicznego w taki sposób, by możliwe było zastosowanie do

ich analizy opisanych dalej metod statystycznych. Warto zwrócić uwagę na ten rozdział. Dobre przygotowanie danych jest kluczem do skutecznego ich opracowywania, i na odwrót — „oszczędzanie czasu" na etapie wstępnego preparowania danych mści się z reguły bardzo pracochłonnym procesem analizy. Treści zawarte w drugim rozdziale uzupełnione są zamieszczonym na końcu skryptu dodatkiem przedstawiającym podstawowe pojęcia z zakresu zmiennych losowych i ich rozkładów. W zasadzie każdy student sięgający po niniejszy skrypt powinien te zagadnienia znać od urodzenia (a w najgorszym wypadku od pierwszego roku studiów), ale wieloletnie doświadczenia autorów wskazują, że pomiędzy tą wiedzą, którą student powinien posiadać, a tą, którą w istocie posiada — bywa spora różnica... Proponujemy więc, aby Czytelnik przeprowadził szybki rachunek sumienia, a w przypadku ujawnienia się wątpliwości — zajrzał do wskazanego dodatku przed przystąpieniem do lektury dalszych rozdziałów skryptu.

Mając zgromadzone i odpowiednio przygotowane dane, można przystąpić do ich analizy. I tu pojawiają się dwie szkoły, wynikające z faktu, że dane biologiczne mają z reguły charakter **wielowymiarowy**. Należy to rozumieć w ten sposób, że dla każdego **pojedynczego** obiektu badań (pacjenta, zwierzęcia doświadczalnego, preparatu tkankowego itp.) rejestruje się **wiele** różnych informacji, które z formalnego punktu widzenia można rozpatrywać jako **wektor**. Składowymi tego wektora mogą być — przykładowo — płeć, wiek, temperatura, morfologia i inne parametry konkretnego pacjenta. Pierwsza szkoła, którą nazwiemy tradycyjną, zakłada analizę poszczególnych tych danych oddzielnie. Można więc określać średni wiek wszystkich pacjentów lub analizować istotność statystyczną podwyższonej temperatury u badanych chorych. Odpowiednie metody biometryczne nazwiemy **jednowymiarowymi** i opiszemy w pierwszej części skryptu — jako prostsze i częściej stosowane. Ponieważ jednak każda zaawansowana analiza statystyczna badanych zjawisk wiąże się w istocie z wykrywaniem zależności między rozpatrywanymi zmiennymi, zatem w drugiej części opisano najważniejsze metody **wielowymiarowe**, to znaczy takie, które uwzględniają fakt istnienia związku występującego pomiędzy zmiennymi, wymagają jednak łącznego (równoczesnego) rozpatrywania tych zmiennych.

Prezentując zawartość skryptu nieco bardziej szczegółowo odnotujemy, że **statystyczny opis pojedynczej zmiennej** otrzymuje się przy użyciu metod opisanych w rozdziale trzecim. Metody te pozwalają na znalezienie pojedynczej **miary**, tzn. liczby, którą można uznać za reprezentantkę wszystkich zgromadzonych obserwacji (tzw. miara tendencji centralnej), a także pozwalają oszacować wielkość przypadkowego rozrzutu obserwowanych danych wokół tej reprezentatywnej miary.

Zamiast mierzyć tendencję centralną i rozrzut lepiej czasem uznać, że opisywane dane wypełniają pewien **przedział** na osi liczbowej. Sposób takiego „przedziałowego" traktowania charakterystyk rozważanych danych opisany jest w rozdziale czwartym.

Same wyniki liczbowe, nawet bardzo wytrawnie opracowane statystycznie, nie są na ogół celem badań. Celem są z reguły pewne **wnioski**, stwierdzające, że coś jest jakieś, na przykład nowy lek jest skuteczniejszy od starego. Aby takie wnioski wyciągnąć

w opraciu o dane obciążone przypadkowym rozrzutem — trzeba uwzględnić ten rozrzut w procesie wnioskowania i trzeba umieć formułować pewne sądy w oparciu o niepewne dane. Techniką używaną w tym celu są tak zwane **testy statystyczne**. W rozdziale piątym opisano najpopularniejsze testy umożliwiające wnioskowanie na podstawie danych **ilościowych** a w rozdziale szóstym — test chi-kwadrat przeznaczony dla danych **jakościowych**.

Uogólnieniem techniki testów statystycznych jest tak zwana **analiza wariancji** opisana w rozdziale siódmym. Ta bardzo pożyteczna technika statystyczna powinna być znacznie częściej stosowana!

Testy zazwyczaj służą do wykrywania **różnic** między zmiennymi, natomiast w biologicznych i medycznych zastosowaniach biometrii równie często poszukujemy **związków** pomiędzy nimi. Rutynową techniką służącą do oceny istnienia (lub braku) tych związków jest obliczanie **korelacji**, zaś metodą matematycznego opisu zachodzących powiązań jest technika **regresji**. Obydwie opisano w rozdziale ósmym. Rozdział ten bezwzględnie powinien być przestudiowany przez każdego studenta nawet w wypadku pominięcia niektórych wcześniejszych rozdziałów!

Testy opisane w rozdziałach 5, 6 i 7 zakładały (niejawnie), że przedmiotem zainteresowania badacza są pewne **parametry** dotyczące rozważanych zmiennych, na przykład wartości średnie w testach Studenta i w analizie wariancji. Czasami zachodzi jednak potrzeba oceny danych bez odwoływania się do jakichkolwiek parametrów — na przykład w celu oceny charakteru rozkładu. W takim wypadku konieczne jest stosowanie **testów nieparametrycznych**, opisanych stosunkowo obszernie w rozdziale dziewiątym. Rozdział ten kończy pierwszą część skryptu, dotyczącą statystyki jednowymiarowej.

Dyskusja **metod wielowymiarowych** rozpoczyna się oddzielnym wprowadzeniem w rozdziale dziesiątym. Warto go uważnie przestudiować, gdyż dostarcza on całościowego spojrzenia na zagadnienia wielowymiarowej analizy danych, które — chociaż trudne — są szczególnie wartościowe w zastosowaniach.

Dyskusja metod wielowymiarowych rozpoczyna się od opisu i analizy **zmiennych wielowymiarowych** (w rozdziale dziesiątym). Na tle tej dyskusji wprowadzona jest (w następnym, jedenastym rozdziale) najpopularniejsza technika wielowymiarowej analizy danych, mianowicie **wielowymiarowa analiza wariancji**. Ta ważna problematyka, uzupełniona **analizą dyskryminacyjną** oraz opisem metod **regresji wielokrotnej** (z uwzględnieniem także w rozdziale dwunastym **regresji nieliowej**, która z obliczeniowego punktu widzenia traktowana musi być jako wielowymiarowa nawet w przypadku zmiennych skalarnych), tworzy zrąb najczęściej stosowanych metod biometrii wielowymiarowej. Skrypt domykają dwa rozdziały tyczące bardziej wyrafinowanych metod **analizy korelacji kanonicznych** w rozdziale czternastym i **analizy czynnikowej** oraz jej odmiany zwanej **analizą głównych składowych** w rozdziale piętnastym.

2. PRZYGOTOWANIE DANYCH

2.1 Rodzaje danych

Badania medyczne lub biologiczne dostarczają danych w postaci poszczególnych obserwacji. Obserwacje te przed ich właściwym wykorzystaniem należy na ogół w jakiś sposób przygotować czy opracować. Niniejszy rozdział poświęcony będzie sposobom wstępnego przygotowania danych. Sposoby przygotowania wstępnego zależne są od rodzaju obserwacji. Rozróżniamy dwa zasadnicze typy obserwacji: jakościowe i ilościowe.

Obserwacje *jakościowe* to takie, które nie mogą być w sposób jednoznaczny i oczywisty scharakteryzowane przy pomocy liczb. Do typowych przykładów należą płeć, grupa krwi, zgon lub przeżycie, obecność lub nieobecność bakterii w badanym preparacie, itd. Opracowanie posiadanych obserwacji jakościowych ma już jednak charakter *ilościowy*, gdyż jest to na ogół zliczenie obserwacji w poszczególnych kategoriach jakościowych.

Przykład 2.1

Pacjentów pewnego szpitala sklasyfikowano według grupy krwi. Uzyskano podział pacjentów na cztery kategorie (klasy): A, B, AB i 0. Liczebności w poszczególnych kategoriach i udziały względne przedstawia tabela 2.1.

Tabela 2.1

Pacjenci pewnego szpitala w rozbiciu względem grupy krwi

Grupa krwi	Liczba	Udział (frakcja)	Udział (w %)
A	425	0,409	40,9%
B	180	0,163	16,3%
AB	84	0,076	7,6%
0	388	0,352	35,2%
Ogółem	1114	1,000	100,0%

W powyższym przykładzie liczba 452 oznacza ilość tych pacjentów, u których stwierdzono grupę A. Liczbę tę (jak i pozostałe z tej kolumny tabeli 2.1) nazywamy *liczebnością* albo *częstością*. Ponieważ wszystkich badanych pacjentów było 1114, więc interesować nas będzie stosunek $\frac{452}{1114} = 0{,}409$ zwany *częstością względną* albo *frakcją* lub czasem *udziałem względnym*. Czasami operujemy także *udziałami procentowymi* (por.ostatnią kolumnę tabeli 2.1)

Tabela 2.2

Liczba dzieci badanych w kierunku nosicielstwa bakterii** Streptococcus pyogenes **w zależności od wielkości migdałków

Stan migdałków	Liczba dzieci	Udział (frakcja)
nie powiększone	516	0,369
powiększone	589	0,421
bardzo powiększone	293	0,210
Ogółem	1398	1,000

Dane przedstawione w tabeli 2.2, podobnie jak dane z tabeli 2.1, są przykładem klasyfikacji jakościowej. Tutaj również żadnej kategorii nie można w sposób absolutnie jednoznaczny scharakteryzować liczbowo, jakkolwiek w odróżnieniu od tabeli 2.1 kategorie mogą być uporządkowane. Porządek może być wprowadzony ze względu na wielkość migdałków w każdej kategorii. Możliwość uporządkowania danych jakościowych pozwala na stosowanie do tych danych bardziej precyzyjnych metod analizy, niż w odniesieniu do danych stricte jakościowych. Z tego względu mówi się niekiedy o skali porządkowej i traktuje się tę kategorię danych jako swoiście *pośrednią* pomiędzy ilościowymi i jakościowymi. Niemniej na danych o ustalonej skali porządkowej, podobnie jak na danych stricte jakościowych nie można wykonywać żadnych operacji arytmetycznych. Jest to bardzo ważne, gdyż wprowadzając dane tego typu do komputera zazwyczaj stosuje się kody liczbowe (na przykład zapisuje się grupę krwi A jako 1, B jako 2, itd.). Zachęca to do stosowania na przykład średniej lub wariancji do opisu większych grup takich danych. Z metodologicznego punktu widzenia jest to jednak niedopuszczalne, jaką bowiem interpretację można w takim wypadku nadać powstałej w wyniku obliczeń wartości — przykładowo — 1,73?

Rozważmy dane z tabeli 2.3. Mimo zewnętrznych podobieństw do poprzednich przypadków, są to już dane o charakterze ilościowym, tyle że zagregowane. Wiek jest zmienną *ilościową*, można go podać np. z dokładnością do dnia, lub jeszcze większą. Na ogół nie

jest to potrzebne, wystarcza dokładność do jednego roku, albo jeszcze dłuższego przedziału czasu, tak jak to ma miejsce w tabeli 2.3.

Tabela 2.3

Rozkład wieku pacjentów z nowotworem płuc w pewnym szpitalu *(według [Armitage])*

Wiek	Liczba pacjentów	Udział (frakcja)
25 ÷ 34	17	0,012
35 ÷ 44	116	0,087
45 ÷ 54	493	0,363
55 ÷ 64	545	0,401
65 ÷ 74	186	0,137
Ogółem	1357	1,000

Niejednokrotnie, zwłaszcza gdy obserwacji nie mamy zbyt wiele, nie możemy zadowolić się takimi zbiorczymi zestawieniami, jakich przykładem jest tabela 2.3 i które nazywamy *szeregiem rozdzielczym*, a musimy uwzględniać dokładne ilościowe wartości wszystkich obserwacji i dopiero na takiej podstawie dokonywać badania statystycznego. Najważniejszą cechą danych jest to, że dane tego typu pozwalają korzystać — bez żadnych ograniczeń — z wszelkich działań arytmetycznych. Zasady przetwarzania danych typu ilościowego są znacznie wygodniejsze i prowadzą do znacznie bardziej precyzyjnych wyników. Warto zauważyć, że dane ilościowe — na przykład poprzez agregację — zawsze można sprowadzić do postaci analogicznej do danych jakościowych i w ten sposób korzystać z metod wyspecjalizowanych do analizy danych tego typu. Podobna zamiana w drugą stronę na ogół jest niemożliwa. (Jakkolwiek istnieją próby tworzenia technik „wzmacniania" skal przy wykorzystaniu informacji pobocznych — por. prace J. Pociechy).

Dane ilościowe dzielimy niejednokrotnie na obserwacje o charakterze ciągłym i dyskretnym. Obserwacje dyskretne mogą przyjmować tylko określone wartości, na przykład może to być pomiar polegający na policzeniu czegoś i wyrażony tylko w liczbach całkowitych (np. liczba dzieci). Obserwacje ciągłe mogą przyjmować w zasadzie dowolne wartości z określonego przedziału. Przykładem takich danych medycznych może być temperatura ciała, ciśnienie krwi, wzrost, ciężar ciała itp.

2.2 Cele przekształcania danych pierwotnych

Wszystkie statystyczne metody opracowywania danych i wnioskowania oparte są na pewnych założeniach, które to założenia nie zawsze są w dostatecznym stopniu spełnione. Piszemy, że założenia nie są spełnione w dostatecznym stopniu, a nie po prostu, że nie są spełnione, gdyż stwierdzenie tego faktu z całkowitą pewnością jest niemożliwe ze względu na brak ostrych kryteriów takiego osądu oraz losowy charakter samych danych. Jednakże często prawdopodobieństwo niezgodności danych z założeniami jest dostatecznie duże, aby postulować potrzebę wstępnego przekształcenia danych pierwotnych x za pomocą pewnej transformacji $f(x)$:

$$y = f(x) \tag{2.1}$$

tak aby uzyskane przekształcone dane charakteryzowały się dużym prawdopodobieństwem zgodności z założeniami określonej metody statystycznej.

Konkretyzując, na ogół dokonujemy przekształcenia (2.1) w jednym z trzech poniższych celów:

1) stabilizacja wariancji,
2) linearyzacja zależności między dwiema cechami,
3) normalizacja rozkładu.

Wariancja jest miarą rozrzutu danych wokół średniej. Często przychodzi nam analizować kilka podgrup danych różniących się średnimi. Znamy bardzo wygodną metodę przeprowadzania badań statystycznych w takich przypadkach, a mianowicie analizę wariancji (por. rozdział 7). Metoda ta wymaga jednak, aby rozrzut wewnątrz poszczególnych podgrup wokół średnich w tych podgrupach był w przybliżeniu jednakowy dla wszystkich podgrup. Jeżeli tak nie jest i miara tego rozrzutu (tzw. wariancja resztowa) jest pewną funkcją wartości średniej w podgrupach

$$\sigma^2(x) = \Phi[E(x)] \tag{2.2}$$

to w celu stabilizacji wariancji można przed przystąpieniem do analizy dokonać takiego wstępnego przekształcenia danych (2.1), aby w przybliżeniu była spełniona zależność:

$$\frac{dy}{dx} = \frac{\text{const}}{\sqrt{\sigma^2(x)}} \tag{2.3}$$

Drugi cel przekształceń wstępnych to linearyzacja zależności między dwiema cechami. Jeżeli dla elementów pewnej zbiorowości będziemy znali wartości dwu zmiennych o charakterze ilościowym (np. dla grupy niemowląt wagę urodzeniową i przyrost wagi między

siedemdziesiątym a setnym dniem po urodzeniu), to możemy badać związek między tymi zmiennymi. Gdy związek ten będziemy mogli w dostatecznie dobrym przybliżeniu przedstawić linią prostą, to wtedy metody analizy będą szczególnie proste, a jej wyniki intuicyjnie zrozumiałe i łatwe do interpretacji. Tak więc niejednokrotnie korzystnie jest zastosować wstępne przekształcenie linearyzujące do danych nie charakteryzujących się pierwotnie zależnością prostoliniową i przeprowadzić badania zwane analizą regresji liniowej już dla danych przekształconych. W pewnych przypadkach (por. podrozdział 2.4) linearyzujące przekształcenia wstępne są koniecznością, z uwagi na ograniczony zakres zmienności jednej ze zmiennych.

Wreszcie cel trzeci — normalizacja rozkładu. Bardzo często wśród założeń wykorzystywanych standardowych metod statystycznych znajdujemy wymóg normalności rozkładu zmiennej w zbiorowości. Można wprawdzie stosować metody nie zakładające rozkładu normalnego, ale metody te są na ogół bardzo złożone i często mniej efektywne. Lepiej więc zastosować wstępną transformację normalizującą i posługiwać się jedną ze standardowych metod. Należy tu zauważyć, że niejednokrotnie to samo przekształcenie równocześnie stabilizuje wariancję, linearyzuje funkcję regresji, jak i normalizuje rozkład. Z drugiej strony pewna ilość metod statystycznych, dotyczących zwłaszcza testów związanych ze średnią, jest mało wrażliwa na pewne odstępstwa od normalności rozkładu. Stąd też w przypadku kolizji celów, wyższy priorytet uzyskują na ogół przekształcenia stabilizujące wariancję lub linearyzujące zależność.

Niektóre częściej używane przekształcenia omówione zostaną w następnym podrozdziale.

2.3 Najczęściej używane przekształcenia danych ilościowych

2.3.1 Przekształcenie logarytmiczne

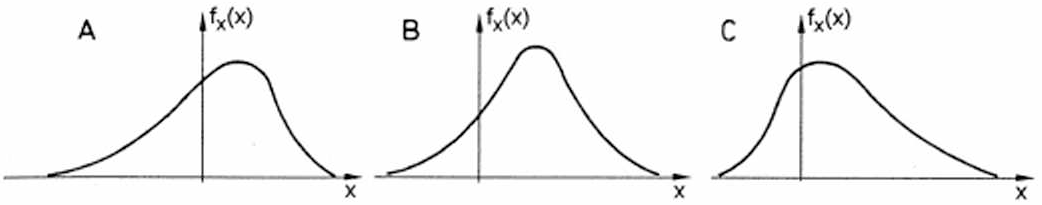
Jeżeli wartość pierwotną obserwacji oznaczymy przez x, a wartość przekształconą przez y, to przekształcenie logarytmiczne wyrazi się zależnością:

$$y = \log x \tag{2.4}$$

Na ogół stosujemy logarytmy dziesiętne lub naturalne. Przekształcać przy pomocy wzoru (2.4) można oczywiście tylko wartości dodatnie. Jeżeli w danych występują zera, można stosować przekształcenie w postaci:

$$y = \log (x + 1) \tag{2.5}$$

Przekształcenie logarytmiczne stosujemy w celu stabilizacji wariancji, gdy w danych pierwotnych $\sigma^2(x)$ rośnie proporcjonalnie do kwadratu x (a właściwie kwadratu wartości oczekiwanej x). Jeżeli związek zmiennej x z jakąś zmienną z charakteryzuje się przebiegiem zbliżonym do wykładniczego (w miarę wzrostu zmiennej z nachylenie linii regresji stale wzrasta), to przekształcenie logarytmiczne będzie taką zależność linearyzować. Przekształcenie logarytmiczne bywa też używane dla normalizacji rozkładów charakteryzujących się asymetrią dodatnią (por. rys. 2.1).



Rys.2.1 Przykładowe rozkłady prawdopodobieństwa ciągłych zmiennych losowych:
A — rozkład asymetryczny (skośny) ujemnie,
B — rozkład symetryczny
C — rozkład asymetryczny (skośny) dodatnio.

Często w badaniach nad efektywnością leków lub szkodliwością trucizn obserwuje się, że średni przyrost efektu ΔE (np. efektu terapeutycznego pewnego leku) jest proporcjonalny do średniego *względnego* przyrostu przyczyny $\frac{\Delta P}{P}$ (ΔP — np. przyrost stężenia leku, czyli dawki leku na jednostkę wagi ciała)

$$\Delta E = k\frac{\Delta P}{P} \tag{2.6}$$

Powyższa zależność po scałkowaniu da nam wzór

$$E = k\ (\ln P) + C \tag{2.7}$$

(k,C — stałe) wskazujący na liniowy związek między efektem a *logarytmem* przyczyny. Tutaj przydatność przekształcenia logarytmicznego nasuwa się sama, zwłaszcza, że w praktycznych badaniach tego typu zwykle stosowane wartości stężenia leku P tworzą tzw. szereg rozcieńczeń, czyli przyjmują wartości ciągu geometrycznego, np. 32, 16, 8, 4, 2, 1 co jest spowodowane kolejnym rozcieńczaniem pierwotnej porcji leku sposobem odlewania połowy roztworu leku i uzupełniania tak powstałego braku wodą. Wielkości tworzące ciąg geometryczny po zlogarytmowaniu dadzą wartości odległe od siebie o stałą wartość.

2.3.2 Przekształcenia pierwiastkowe i kwadratowe

Przekształcenia pierwiastkowe i kwadratowe należą do przekształceń potęgowych danych ogólną zależnością:

$$y = x^a \tag{2.8}$$

W przekształceniu pierwiastkowym $a = \frac{1}{2}$, zaś w kwadratowym $a = 2$. O innym przekształceniu potęgowym, a mianowicie o przekształceniu odwrotnościowym ($a = -1$) będzie traktował następny punkt.

Przekształcenie pierwiastkowe

$$y = \sqrt{x} \tag{2.9}$$

stabilizuje wariancję, gdy jest ona proporcjonalna do średniej (albo równa jej — jak to jest w przypadku rozkładu Poissona). Przy pomocy tego przekształcenia można linearyzować związki charakteryzujące się przebiegiem zbliżonym do kwadratowego. Przekształcenie to bywa także używane do normalizacji rozkładów skośnych dodatnio.

Przekształcenie pierwiastkowe może być używane do danych mikrobiologicznych w postaci zliczeń. W celu oszacowania stężenia drobnoustrojów w zawiesinie, działa się w sposób następujący: najpierw rozcieńcza się badaną zawiesinę np. w stosunku $1:10^5$. Następnie pobiera się z rozcieńczonej zawiesiny próbki o stałej objętości, np. 1 cm^3 i umieszcza się w naczyniach z pożywką. Po pewnym czasie z każdego drobnoustroju, który znalazł się na pożywce rozwinie się kolonia. Średnia liczba kolonii w naczyniu pozwala oszacować ilość drobnoustrojów w 1 cm^3 rozcieńczonej zawiesiny, czyli w 10^{-5} cm^3 zawiesiny pierwotnej. Liczba kolonii w naczyniu jest wielkością podlegającą rozkładowi Poissona i do tego typu danych może być stosowane przekształcenie pierwiastkowe.

Przekształcenie pierwiastkowe bywa także używane do danych w postaci częstości względnych (frakcji), jeżeli są one zawarte w przedziale $0 \div 0{,}2$ lub $0{,}8 \div 1$ (w tym ostatnim przypadku obliczamy najpierw $p' = 1 - p$ i dopiero potem stosujemy przekształcenie). Gdy zakres rozpatrywanych frakcji jest szerszy, a w szczególności rozciąga się od 0 do 1, wtedy stosujemy specjalne przekształcenia, o których będzie mowa w podrozdziale 2.4.

Przekształcenie kwadratowe

$$y = x^2 \tag{2.10}$$

stabilizuje wariancję, gdy wariancja x zmniejsza się w miarę wzrostu średniej, linearyzuje zależność krzywoliniową, gdy funkcja regresji wykazuje zmniejszającą się co do wartości

bezwzględnej pochodną oraz normalizuje rozkłady skośne ujemnie. Przekształcenie to nie jest tak często stosowane, jak przekształcenia omawiane poprzednio.

2.3.3 Przekształcenie odwrotnościowe

Przekształcenie odwrotnościowe należy również do grupy przekształceń potęgowych. Jest ono określone wzorem:

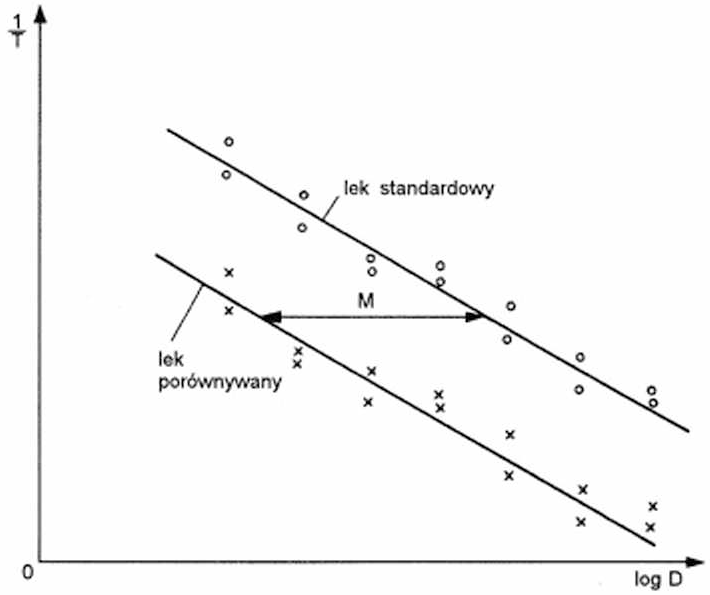
$$y = \frac{1}{x} = x^{-1} \tag{2.11}$$

Przekształcenie to stabilizuje wariancję, gdy jest ona proporcjonalna do czwartej potęgi średniej. Jednakże inna cecha tego przekształcenia powoduje, że jest ono często stosowane. Chodzi mianowicie o to, że dużym wartościom x przypisuje ono praktycznie zerowe wartości przekształcone, zaś niewielkim wartościom x odpowiadają stosunkowo duże wartości y. Przekształcenie takie ułatwia analizę danych w postaci czasów przeżycia zwierząt doświadczalnych uzyskiwanych np. podczas badań efektywności leków. Zastosowanie tego przekształcenia umożliwia linearyzację spotykanych w tym postępowaniu zależności. Szczegóły omówiono w poniższym przykładzie.

Przykład 2.2

Dla porównania efektu terapeutycznego dwóch leków, zakażono dwie grupy zwierząt doświadczalnych tysiąckrotną „pięćdziesięcioprocentową dawką śmiertelną" pewnego wirusa. Sposób ustalania pięćdziesięcioprocentowej dawki śmiertelnej, czyli takiego stężenia wirusa, przy którym połowa zwierząt doświadczalnych ginie, będzie opisany w następnym podrozdziale. Teraz zakażone zwierzęta leczono badanymi lekami podając jednej grupie jeden z leków, a drugiej — drugi. W ramach każdej z grup aplikowano poszczególnym zwierzętom różne dawki, a właściwie różne stężenia odpowiedniego leku. Rejestrowano czasy przeżycia poszczególnych zwierząt, rodzaj podawanego im leku oraz jego stężenie podczas terapii. Następnie tak uzyskane dane poddawano przekształceniom wykorzystując logarytmy stężenia leków (log D) oraz odwrotności czasów przeżycia ($1/T$). Przekształcone dane naniesiono na układ współrzędnych o osiach log D oraz 1/T (patrz rys. 2.2). Dalsze badania statystyczne wykazały, że z dostatecznie dużym prawdopodobieństwem można:

1) zależności między $1/T$ a log D dla każdego z dwóch badanych leków traktować jako prostoliniowe o współczynnikach kierunkowych prostych różnych od zera (por. podrozdział 8.3),
2) obie proste obrazujące te zależności uważać za równoległe (por. punkt 8.4.1),
3) traktować równoległe proste dla obu leków jako proste nie pokrywające się i oszacować ich poziomą odległość M (por. punkty 8.4.3, 8.4.5, a także rys. 2.2).



Rys. 2.2 Przykładowe dane z badania efektu terapeutycznego dwóch leków wraz z dopasowanymi równoległymi prostymi regresji. Kółeczkami oznaczono dane dotyczące leku pierwszego, krzyżykami — leku drugiego.

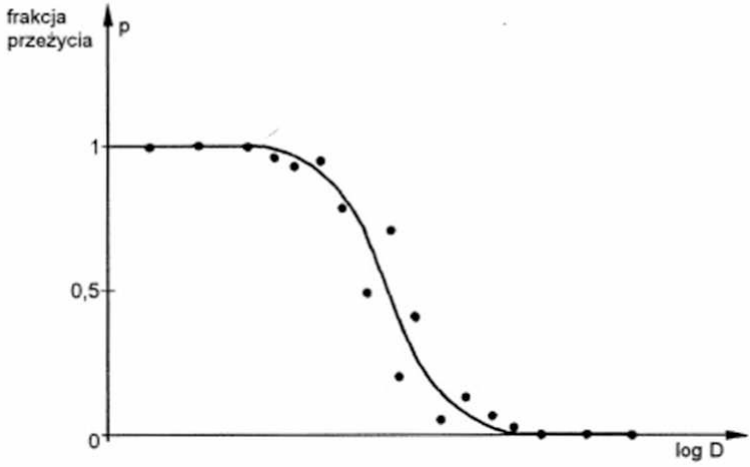
Bez zastosowania odpowiednich przekształceń wstępnych uzyskanie takich wyników nie byłoby możliwe. Pozioma odległość M równoległych prostych obrazujących zależność efektu terapeutycznego (chodzi tu o przekształcony czas przeżycia) od stężenia leku (a właściwie od logarytmu jego stężenia) jest nazywana logarytmem stosunku mocy. Zwykle jeden z leków (np. pierwszy w naszym przykładzie — por. rys. 2.2) traktuje się jako standardowy i służy on jako baza porównawcza dla leku drugiego — testowanego. M jest równe różnicy logarytmów stężeń wywołujących ten sam efekt terapeutyczny, czyli logarytmowi stosunku tych stężeń. Po antylogarytmowaniu wielkości M uzyskamy więc liczbę mówiącą, ile razy większe stężenie leku standardowego w stosunku do stężenia leku badanego jest konieczne dla osiągnięcia tego samego efektu. Stanowi to pewną miarę efektywności leku badanego, który (jak to ma miejsce w naszym przypadku pokazanym na rys. 2.2) wymaga mniejszej dawki dla tego samego efektu, czyli daje lepszy efekt przy tej samej dawce.

2.4 Przekształcenia frakcji

W badaniach skuteczności leków bądź w ocenie szkodliwości środków trujących często uzyskujemy dane w postaci frakcji, przy czym frakcje te wypełniają cały przedział od 0 do 1.

Przykład 2.3

Dokonano próby wyznaczenia *pięćdziesięcioprocentowej dawki śmiertelnej* pewnej trucizny. W tym celu dużą liczbę zwierząt doświadczalnych podzielono na równoliczne grupy i każdej grupie podano truciznę w innym stężeniu. Po upływie określonego czasu oznaczono w każdej grupie frakcję zwierząt, które przeżyły. Dla małych stężeń trucizny frakcje te wynosiły 1, dla bardzo dużych 0. Przy stężeniach pośrednich frakcje były różne od wartości krańcowych i układały się (przy zastosowaniu logarytmicznego przekształcenia stężeń) w przybliżeniu zgodnie z tzw. krzywą sigmoidalną (por. rys. 2.3).



Rys. 2.3 Przykładowe wyniki badań skuteczności pewnej trucizny — dane w postaci frakcji wraz z dopasowaną sigmoidalną krzywą regresji.

Dalsze badanie uzyskanych danych można przeprowadzić po wstępnym przekształceniu frakcji, mającym na celu linearyzację zależności, zwłaszcza w tym zakresie gdzie frakcje przyjmują wartości różne od 0 i 1. Obydwa zakresy prostoliniowego (płaskiego) przebiegu zależności frakcji od stężenia nie niosą żadnej istotnej informacji, gdyż śledząc poziomy przebieg wykresu (dla $p = 1$ i dla $p = 0$) nie wiemy, jak daleko znajdujemy się od strefy zmienności. Dlatego dane te najczęściej eliminuje się z rozważań. Dalszym celem przekształcenia powinna być stabilizacja rozrzutu. Rozrzut jest zerowy w obydwu poziomych prostoliniowych fragmentach wykresu linii regresji, zaś maksymalny w okolicach frakcji $p = 0{,}5$. Przy zbliżaniu się do krańcowych wartości frakcji rozrzut maleje, a jednocześnie rozkład wartości frakcji staje się wyraźnie asymetryczny ze skośnością w *kierunku środka* tj. punktu $p = 0{,}5$ co jest spowodowane naturalnym ograniczeniem zmienności frakcji do przedziału <0, 1>.

Stosuje się trzy przekształcenia danych w postaci frakcji. Trudno stwierdzić, czy któreś z nich jest wyraźnie lepsze od innych. Przekształcenia lepiej realizujące linearyzację nie stabilizują wariancji, zaś stabilizujące wariancję nie w pełni linearyzują

krzywą sigmoidalną. Po krótkim omówieniu wspomnianych przekształceń powrócimy do przykładu 2.3.

2.4.1 Przekształcenie kątowe

Przekształcenie kątowe dane zależnością

$$y = \arcsin \sqrt{p} \tag{2.12}$$

stabilizuje wariancję, która jest równa

$$\sigma^2 (y) \cong \frac{820{,}7}{n} \tag{2.13}$$

(gdzie n — liczebność próby), jeżeli rozkład jest ściśle dwumianowy. Natomiast linearyzacja zależności sigmoidalnych przy przekształceniu tym nie jest idealna: obserwuje się pewne spłaszczenia wykresu przy górnej i dolnej granicy przedziału zmienności (tzn. dla $p \approx 0$ i $p \approx 1$), choć w większej części przedziału krzywa sigmoidalna jest dobrze „prostowana". Tabela 2.4 zawiera krótką tablicę przekształcenia kątowego, jak i dwóch dalszych omawianych poniżej.

Tabela 2.4

Krótka tablica przekształceń frackcji *(według [Armitage])*

Frakcja	Przekształcenie		
	kątowe (w stopniach kątowych)	logitowe	probitowe
0	0	$-\infty$	$-\infty$
0,05	13	–2,94	3,36
0,10	18	–2,20	3,72
0,15	23	–1,73	3,96
0,20	27	–1,39	4,16
0,25	30	–1,10	4,33
0,30	33	–0,85	4,48
0,35	36	–0,62	4,61
0,40	39	–0,41	4,75
0,45	42	–0,20	4,87
0,50	45	0	5,00

Frakcja	Przekształcenie		
	kątowe (w stopniach kątowych)	logitowe	probitowe
0,55	48	0,20	5,13
0,60	51	0,41	5,25
0,65	54	0,62	5,39
0,70	57	0,85	5,52
0,75	60	1,10	5,67
0,80	63	1,39	5,84
0,85	67	1,73	6,04
0,90	72	2,20	6,28
0,95	77	2,94	6,64
1,00	90	∞	∞

2.4.2 Przekształcenie logitowe

Logit y frakcji p definiujemy jako

$$y = \ln \frac{p}{1-p} \tag{2.14}$$

W przekształceniu tym pomijamy obserwacje, dla których $p = 0$ lub $p = 1$. Obserwacje te, jak wiemy, nie wnoszą istotnej informacji, a odpowiadające im wartości przekształcone są nieskończone (nie istnieją). Czasami zamiast pomijać wartości krańcowe $p = 0$ i $p = 1$ modyfikuje się nieco wzór (2.10) tak, że przekształcenie logitowe definiuje się jako

$$y = \ln \frac{r + \frac{1}{2}}{n - r + \frac{1}{2}} \tag{2.15}$$

gdzie $p = \frac{r}{n}$, natomiast n jest liczebnością próby wykorzystywanej do wyznaczenia frakcji.

Przekształcenie logitowe jest podobne do kątowego w takim rozumieniu, że także bardziej „rozciąga" końce skali p niż jej środek. Jednakże w bezpośredniej bliskości wartości granicznych podobieństwo znika, gdyż przekształcenie logitowe jeszcze bardziej rozciąga skalę przyjmując wartości dowolnie duże (skala przekształcenia kątowego jest

ograniczona). Przekształcenie logitowe dobrze linearyzuje krzywe sigmoidalne, nie zapewniając jednak stabilizacji wariancji.

2.4.3 Przekształcenie probitowe

Przekształcenie to, jakkolwiek z trzech tutaj omawianych najbardziej skomplikowane, jeżeli chodzi o wyartykułowanie i sprawiające trudności obliczeniowe (wymaga dysponowania tablicami statystycznymi), jest najczęściej używane, co prawdopodobnie ma swoje podłoże w głęboko zakorzenionej wśród przedstawicieli nauk medycznych i biologicznych tradycji.

Niech dla dowolnej frakcji p liczba y' będzie taką wartością, że na lewo od y' znajduje się p-ta część powierzchni zawartej pod krzywą standaryzowanego rozkładu normalnego. Innymi słowy y' można określić z zależności

$$p = F(y') \tag{2.16}$$

gdzie $F(\cdot)$ jest dystrybuantą standaryzowanego rozkładu normalnego (rozkładu normalnego o wartości średniej równej zeru i odchyleniu standardowym równym jeden).

Probit y frakcji p jest definiowany jako:

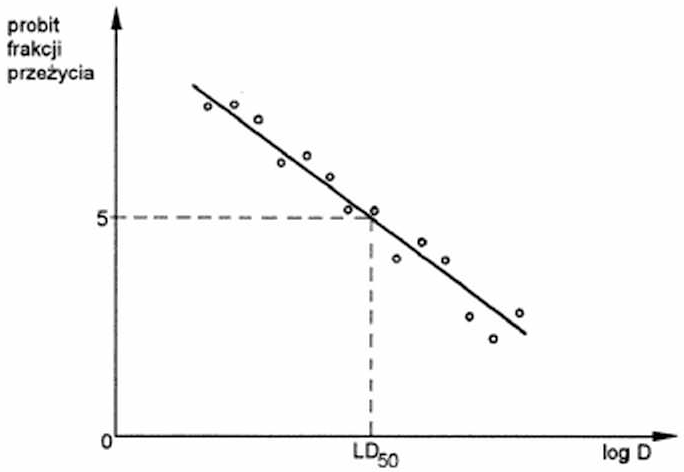
$$y = 5 + y' \tag{2.17}$$

Dodanie wartości 5 do y' w powyższej definicji wynika z tradycji i nie ma jakiegoś merytorycznego uzasadnienia.

Przekształcenie probitowe ma własności bardzo podobne do przekształcenia logitowego. Również przyjmuje wartości nieskończone dla $p = 0$ i $p = 1$, dobrze linearyzuje regresję sigmoidalną i nie stabilizuje wariancji. Niekiedy, aby uniknąć trudności z nieskończonymi wartościami y dla $p = 0$ i $p = 1$ przyjmuje się poprawkę w definicji $p = \frac{r}{n}$ określając: dla $r = 0$ (czyli $p = 0$) $p' = \frac{1}{2n}$ i dla $r = n$ (czyli $p = 1$) $p' = \frac{2n - 1}{2n}$ i później w granicznych przypadkach używa się wartości p' przy obliczaniu probitów.

Przykład 2.3 (ciąg dalszy)

Kontynuujmy przykład poświęcony wyznaczaniu *pięćdziesięcioprocentowej dawki śmiertelnej* pewnej trucizny. Dane w postaci frakcji przeżycia dla różnych logarytmów stężeń trucizny przekształcamy obliczając ich probity i zaznaczając to na wykresie (por. rys. 2.4). Następnie wykorzystując odpowiednie metody analizy regresji (opisane dalej w rozdziale 8) dopasowujemy do uzyskanych punktów w najlepszy możliwie sposób prostą regresji, będącą obrazem zależności probitu frakcji przeżycia od logarytmu



Rsy. 2.4 Ilustracja sposobu określania „50%-owej dawki śmiertelnej" LD_{50} pewnej trucizny z wykorzystaniem przekształcenia probitowego.

stężenia. Ponieważ frakcji $p = 0{,}5$ (czyli 50%) odpowiada probit równy 5, więc prosta równoległa do osi odciętych przechodząca przez punkt (0, 5) da na przecięciu z prostą regresji punkt, którego odcięta będzie właśnie logarytmem szukanej 50% — owej dawki śmiertelnej. Wartość tę (po antylogarytmowaniu) używamy do badań takich, jak opisane w przykładzie 2.2.

2.5 Eliminowanie obserwacji nietypowych

Podczas wstępnej analizy danych niejednokrotnie spotykamy się z obserwacjami, które wydają się niewiarygodne lub znacznie odbiegają od pozostałych wyników. Powody takiego stanu rzeczy mogą być dwa. Pierwszym są różnego rodzaju błędy: wynikające z niepoprawnej metody badawczej, złego przeprowadzenia eksperymentu, złych warunków prowadzenia doświadczeń, pomyłek w zapisywaniu wyników na papierze, pomyłek w przenoszeniu danych na nośniki komputerowe itd. Drugim powodem jest tzw. zmienność losowa materiału statystycznego. Przykładem może być tutaj natrafienie podczas wstępnej analizy danych szpitalnych na kartę chorego, którego wzrost wynosi 222 cm. Jest to wartość budząca podejrzenia, jednak nie jest to wartość zupełnie nieprawdopodobna. Jeżeli może ona być w jakiś sposób potwierdzona, należy ją oczywiście przyjąć do dalszych obliczeń. Odrzucenie byłoby tu błędem sztuki. Natomiast wzrost dorosłego pacjenta równy 272 cm lub 27 cm należy odrzucić z powodów oczywistych — jako błąd gruby.

Ostatni przykład jest ilustracją jednej z metod eliminacji obserwacji nietypowych, a mianowicie sprawdzania logicznego. Polega ono na eliminowaniu pewnych wartości,

które można uważać za niemożliwe lub krańcowo nieprawdopodobne z racji znaczenia samej obserwacji. Przykładem może być temperatura ciała wynosząca 10°C lub podanie w rubryce „stan cywilny" dziesięcioletniej pacjentki informacji „zamężna". Tego typu błędne dane mogą być eliminowane ręcznie bądź też automatycznie z wykorzystaniem komputera, po ustaleniu odpowiednich dla każdego typu danych reguł i zasad.

Drugim sposobem eliminacji jest wykorzystanie znajomości własności statystycznych materiału obserwacyjnego. Najbardziej znaną zasadą z tej grupy jest tzw. *reguła trzech sigm*. Otóż najczęściej spotykamy się z danymi o rozkładzie normalnym lub zbliżonym do normalnego. W takich przypadkach prawdopodobieństwo tego, że wartość zmiennej losowej znajdzie się w przedziale, którego środkiem jest wartość oczekiwana μ, a granice wynoszą $\mu - 3\sigma$ oraz $\mu + 3\sigma$ (σ to oznaczenie odchylenia standardowego), jest równe 0,9973 czyli jest to zdarzenie prawie pewne. Jeżeli więc zaobserwujemy wartości spoza przedziału o promieniu 3σ, to do tej wartości należy podejść z dużym sceptycyzmem. Postulowane jest wyeliminowanie jej, zwłaszcza jeżeli istnieje jakiś dodatkowy powód zewnętrzny wskazujący na przykład, że obserwację wykonał niedoświadczony technik lub że agregat klimatyzacyjny uległ awarii w pewnym momencie trwania eksperymentu, itd. Gdy jednak nie podejrzewa się wpływu dodatkowych czynników zewnętrznych, to wskazane jest raczej pozostawienie podejrzanej obserwacji.

W ogóle należy stwierdzić, że problem eliminacji obserwacji nietypowych i błędnych jest zagadnieniem bardzo trudnym i ciężko poddającym się jakiejkolwiek unifikacji. Wymaga dużej wnikliwości, doświadczenia i rutyny oraz indywidualnego podejścia do każdego przypadku. Jednocześnie jest to zagadnienie bardzo ważne, gdyż decyzja odnośnie pozostawienia lub wyeliminowania obserwacji krańcowych może w znacznym stopniu rzutować na wyniki całej analizy statystycznej.

3. STATYSTYKA OPISOWA

Statystyka zajmuje się metodami wnioskowania o całej *zbiorowości statystycznej* (zwanej czasami populacją generalną) na podstawie zbadania pewnej jej części zwanej *próbą*. Metody wnioskowania statystycznego w zastosowaniu do problemów biologii i medycyny będą tematem dalszych rozdziałów. W szczególności w rozdziale czwartym będzie się mówić o *estymacji*, czyli szacowaniu parametrów rozkładu badanej cechy w populacji generalnej na podstawie znajomości wyników próby. Dalsze rozdziały podręcznika poświęcone będą *weryfikacji hipotez* statystycznych, dotyczących rozkładu badanej cechy w zbiorowości generalnej. Testowanie tych hipotez także będzie brało za podstawę wyniki próby pobranej z populacji generalnej.

Niniejszy rozdział nie będzie jeszcze zajmował się właściwym wnioskowaniem statystycznym. Poświęcimy go wstępnemu badaniu wyników próby. Istnieje często potrzeba wyrażenia serii wartości (np. pomiarów) w postaci jednej liczby odzwierciedlającej ogólny poziom zjawiska, jego przeciętną tendencję. Liczby tego typu nazywane bywają miarami skupienia, miarami położenia lub chyba najwłaściwiej miarami tendencji centralnej. Najważniejszą z nich jest **średnia**. Poza wskazaniem liczby, wokół której koncentrują się wyniki próby niezbędne jest niejednokrotnie określenie stopnia rozproszenia wyników próby wokół średniej. Jest to ważne np. dla oceny wiarygodności oszacowania (estymacji) średniej. Jeśli chcemy cokolwiek wnioskować o średniej w całej populacji na podstawie znajomości średniej z próby, to w przypadku gdy poszczególne wyniki próby mało różnią się wzajemnie — podchodzimy do naszego szacunku ze znacznie większym zaufaniem niż wówczas, gdy rozrzut wyników próby jest duży. Dla zbadania stopnia rozproszenia danych stosujemy różne miary rozrzutu (miary zmienności, miary dyspersji), z których najważniejszą jest **wariancja**. Poniżej podamy pewne podstawowe wiadomości dotyczące miar tendencji centralnej i miar rozrzutu. Informacje te należą do dziedziny zwanej statystyką opisową.

3.1 Miary tendencji centralnej

3.1.1 Średnia, mediana, wartość modalna

Najczęściej używaną miarą tendencji centralnej jest **średnia** z próby. Oznaczamy ją $\bar{x}$ i definiujemy jako średnią arytmetyczną wyników próby:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{3.1}$$

przy czym:

x_i — obserwacja wartości badanej cechy dla i-tego elementu populacji generalnej wybranego do próby,

n — ilość wszystkich obserwacji w próbie.

Innym miernikiem tendencji centralnej jest **mediana**. Jest to wartość obserwacji środkowej, jeżeli wcześniej uporządkowaliśmy wszystkie obserwacje w kolejności np. rosnących wartości. Gdy liczba obserwacji n jest nieparzysta — medianą jest obserwacja o numerze $\frac{1}{2}(n+1)$. Jeżeli mamy parzystą liczbę obserwacji to przyjmujemy, że medianą jest średnia dwóch obserwacji środkowych, to znaczy obserwacji o numerze $\frac{1}{2}n$ oraz obserwacji stojącej na miejscu $\frac{1}{2}n + 1$.

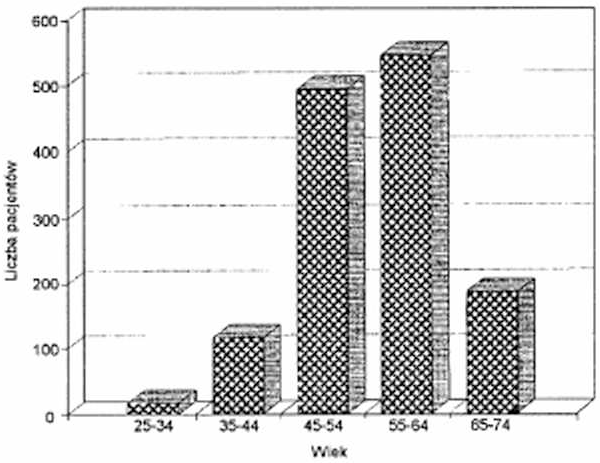
Często dokonuje się porównań obu miar tendencji centralnej. Zwraca się wówczas uwagę na następujące zagadnienia:

1) obliczając średnią korzystamy z wyników wszystkich obserwacji, mediana jest zaś pojedynczą obserwacją lub zależy od co najwyżej dwu obserwacji. Dlatego też mediana niesie w sobie mniej informacji o próbie niż średnia,
2) zmiany wartości obserwacji ekstremalnych nie mają wpływu na wielkość mediany, a wpływają na średnią. Z tego powodu dla silnie asymetrycznych (skośnych — por. rys. 2.1) rozkładów obserwacji mediana jest lepszym miernikiem tendencji centralnej, gdyż lepiej odzwierciedla typowe wartości obserwacji,
3) mediana w niewielkim stopniu nadaje się do przekształceń i obliczeń matematycznych, nie jest więc zbyt często wykorzystywana w zaawansowanych metodach statystycznych.

Kolejną miarą tendencji centralnej jest **wartość modalna** (dominanta, moda). Określa się ją jako wartość tej (tych) obserwacji, która występuje najczęściej w danej próbie. Miernik ten charakteryzuje się dużą zmiennością w próbach o niewielkiej liczbie obserwacji. Jest on rzadko używany w analizie statystycznej.

3.1.2 Obliczanie średniej, mediany i modalnej dla danych w postaci szeregów rozdzielczych

W poprzednim rozdziale w tabeli 2.3 pokazano przykład danych ilościowych w formie tzw. szeregu rozdzielczego. Taki zagregowany sposób dostarczania danych stosuje się w przypadku dużej liczby obserwacji. Grupuje się wówczas obserwacje w kilka do kilkunastu klas oraz podaje się jedynie granice przedziałów klasowych i liczby obserwacji w poszczególnych klasach. Dane takie często przedstawia się graficznie w postaci histogramów (por. rys. 3.1).



Rys. 3.1 Histogram rozkładu wieku pacjentów z nowotworem płuc (według tabeli 2.3).

Niejednokrotnie istnieje potrzeba obliczenia wartości średniej, mediany i wartości modalnej dla danych w postaci szeregu rozdzielczego. Stosujemy wówczas wzory przybliżone, które przedstawimy poniżej. Wzory te wykorzystamy dalej do obliczeń mierników tendencji centralnej danych dotyczących wielu pacjentów z nowotworem płuc (por. tabela 2.3). Dane te powtórzono w tabeli 3.1 uzupełniając je dodatkowo o postać szeregu skumulowanego (dla danych przedziału klasowego podaje się sumę liczebności danej klasy i wszystkich poprzednich — jest to pewien odpowiednik dystrybuanty rozkładu).

Wartość średnią dla danych w postaci szeregu rozdzielczego oblicza się według wzoru:

$$\overline{x} \cong \frac{\sum_{i=1}^{k} \overset{\circ}{x}_i n_i}{\sum_{i=1}^{k} n_i} \tag{3.2}$$

gdzie:

n_i — liczebność w i-tym przedziale klasowym,

k — liczba klas,
$\overset{\circ}{x}_i$ — środek i-tego przedziału klasowego.

Tabela 3.1

Dane do przykładu obliczania miar tendencji centralnej

Wiek		Liczba pacjentów		
Przedziały klasowe (granice)	Środki przedziałów klasowych	Szereg rozdzielczy	Szereg skumulowany	Wartości pomocnicze do obliczania średniej
	$\overset{\circ}{x}_i$	n_i		$n_i \cdot \overset{\circ}{x}_i$
25 — 34	30	17	17	510
35 — 44	40	116	133	4640
45 — 54	50	493	626	24650
55 — 64	60	545	1171	32700
65 — 74	70	186	1357	13020
		1357		75520

W rozważanym przypadku (tab. 3.1)

$$\bar{x} = 75520/1357 = 55{,}6 \text{ lat}$$

Medianę wyznacza się w następujący sposób:

1) określa się numer obserwacji, której wartość jest medianą (w naszym przykładzie $n = 1357$ jest nieparzyste, więc

$$N_{Me} = (n + 1)/2 = 1358/2 = 679\ ),$$

2) na podstawie szeregu skumulowanego odnajduje się klasę w której leży mediana (u nas klasa 55 — 64 lat),
3) oblicza się wartość mediany *Me* z następującego wzoru wykorzystującego metodę interpolacji liniowej:

$$Me = x_0 + \frac{l}{n_0}\left(N_{Me} - N^*\right) \tag{3.3}$$

gdzie:

x_0 — dolna granica przedziału klasowego mediany,
l — rozpiętość przedziału klasowego mediany,
n_0 — liczebność w przedziale mediany,
N_{Me} — numer obserwacji, której wartość jest medianą (= n / 2 lub (n + 1) / 2),
N^* — skumulowana liczba obserwacji do klasy mediany.

W naszym przypadku otrzymuje się:

$$Me = 55 + (10/545) * (679 - 626) = 56 \text{ lat} .$$

Aby obliczyć wartość modalną ustala się przedział klasowy modalnej (jest to ten przedział, w którym liczebność jest największa — w naszym przykładzie 55 — 64 lat). Następnie oblicza się wartość modalną D ze wzoru:

$$D = x_0 + l\frac{n_d - n_{d-1}}{(n_d - n_{d-1}) + (n_d - n_{d+1})} \tag{3.4}$$

gdzie:

x_0 — dolna granica przedziału klasowego modalnej,
l — rozpiętość przedziału modalnej,
n_d — liczebność w przedziale modalnej,
n_{d-1} — liczebność w przedziale poprzedzającym klasę modalnej,
n_{d+1} — liczebność w przedziale następującym po klasie modalnej,

W rozważanym przykładzie

$$D = 55 + 10\frac{545 - 493}{(545 - 493)\,(545 - 186)} = 56{,}3 \text{ lat}$$

Uzyskaliśmy wartości miar tendencji centralnej spełniające zależność

modalna > mediana > średnia

typową dla rozkładów ujemnie skośnych (por. rys. 3.1).

3.1.3 Średnia geometryczna i średnia harmoniczna

W podpunktach 2.3.1 oraz 2.3.3 omówiono przekształcenie logarytmiczne i przekształcenie odwrotnościowe. Z przekształceniami tymi związane są pojęcia średnich: geometrycznej i harmonicznej. Jeżeli bowiem przed dokonaniem analizy podda się dane przekształceniu logarytmicznemu:

$$y = \log_a x \tag{3.5}$$

a następnie wyliczy się średnią arytmetyczną $\bar{y}$ danych przekształconych, to wartość ta po powrocie do pierwotnej skali danych (po antylogarytmowaniu)

$$\bar{x}_g = a^{\bar{y}} \tag{3.6}$$

da wielkość $\bar{x}_g$ będącą średnią geometryczną danych pierwotnych. Często przytaczany w literaturze wzór dla średniej geometrycznej

$$\bar{x}_g = \sqrt{x_1 \cdot x_2} \tag{3.7}$$

jest oczywiście szczególnym przypadkiem w omówionej wyżej procedurze, gdyż niezależnie od logarytmicznej „techniki" wyliczania średniej geometrycznej, jej wartość w ogólnym przypadku wyznaczana jest zgodnie z wzorem:

$$\bar{x}_g = \left(\prod_{i=1}^{n} x_i \right)^{\frac{1}{n}} \tag{3.8}$$

gdzie symbol Π użyto do oznaczenia iloczynu wielu argumentów analogicznie do powszechnie znanego symbolu Σ.

Podobnie jeżeli dokona się przekształcenia danych z próby przez odwrotność

$$y = 1/x \tag{3.9}$$

i wyliczy się średnią arytmetyczną $\bar{y}$ danych przekształconych, to po przejściu do poprzedniej skali danych

$$\bar{x}_h = 1/y \tag{3.10}$$

uzyska się wielkość x_h będącą średnią harmoniczną danych pierwotnych. Ogólny wzór dla średniej harmonicznej jest następujący:

$$\bar{x}_h = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} \tag{3.11}$$

Obydwie średnie: geometryczna i harmoniczna są mniejsze lub równe średniej arytmetycznej, przy czym równość zachodzi tylko dla identycznych wszystkich uśrednianych wartości.

Obie rozważane średnie nie są na ogół stosowane zamiast średniej arytmetycznej jako alternatywne miary tendencji centralnej w kompleksowej analizie statystycznej. Stanowią one niejako efekt uboczny wykorzystywania wstępnych przekształceń danych w problemach statystycznych w biologii i medycynie (por. rozdział 2).

3.2 Miary rozrzutu

3.2.1 Rozstęp, odchylenie ćwiartkowe, odchylenie przeciętne

Najprostszą miarą rozrzutu jest rozstęp, czyli różnica pomiędzy największą i najmniejszą obserwacją. Jest to bardzo naturalny miernik rozrzutu, wygodny zwłaszcza wówczas, gdy trzeba szybko otrzymać orientacyjną charakterystykę zmienności danej próby. Prostota definicji rozstępu jest jednak także jego wadą. Wartość rozstępu zależy bowiem tylko od dwóch obserwacji i w dodatku są to obserwacje skrajne, a wnioskowanie oparte wyłącznie na wartościach krańcowych nie jest wiarygodne. Gubiona jest wówczas cała informacja o zmienności obserwacji położonych wewnątrz przedziału ograniczonego wartościami ekstremalnymi, a zmienność ta może być mocno zróżnicowana. Poza tym rozstęp wykazuje dużą zmienność przy zmianach próby. Na ogół rośnie on przy wzroście liczebności próby, gdyż do dużych prób łatwiej zostaną wylosowane obserwacje nietypowe, rzadko pojawiające się w populacji i wyraźnie różniące się od pozostałych. Z tych to powodów nie wykorzystuje się szerzej rozstępu jako miary rozrzutu.

Wymienione powyżej wady w mniejszym stopniu dotyczą następnej z miar rozrzutu, a mianowicie odchylenia ćwiartkowego. Odchylenie ćwiartkowe to miernik oparty także na wartościach dwóch obserwacji, ale już nie obserwacji skrajnych. Chodzi o tzw. kwartyle. Wartość obserwacji, poniżej której znajduje się $\frac{1}{4}$ uporządkowanych obserwacji, nazywa się kwartylem dolnym, zaś wartość obserwacji, powyżej której znajduje się $\frac{1}{4}$ uporządkowanych obserwacji, to kwartyl górny.

Różnicę między obydwoma kwartylami nazywamy odchyleniem ćwiartkowym. Miernik ten, mimo że jest znacznie bardziej stabilny niż rozstęp, także nie jest zbyt często używany z uwagi m.in. na trudności w określeniu numerów obserwacji będących kwartylami — zwłaszcza dla małych prób.

Na informacjach o odchyleniach wartości wszystkich obserwacji od średniej bazuje następna miara rozrzutu, a mianowicie odchylenie przeciętne zwane niekiedy odchyleniem średnim. Jest ono zdefiniowane wzorem

$$d = \frac{\sum_{i=1}^{n} \left| x_i - \bar{x} \right|}{n} \tag{3.12}$$

jako średnia arytmetyczna wartości bezwzględnych odchyleń poszczególnych obserwacji od średniej. Miara ma tę wadę, że trudno poddaje się działaniom matematycznym (obecność wartości bezwzględnych) i nie posiada tak bezpośredniej interpretacji teoretycznej jak omawiane dalej odchylenie standardowe - więc także i ona nie ma szerszego zastosowania w statystyce.

3.2.2 Wariancja i odchylenie standardowe. Problem estymacji punktowej

Wariancję definiujemy wzorem

$$\sigma^2 = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2}{n} \tag{3.13}$$

Jest to średnia arytmetyczna sumy kwadratów odchyleń wartości obserwacji od średniej. Pierwiastek kwadratowy wariancji nazywamy odchyleniem standardowym

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2}{n}} \tag{3.14}$$

Wariancja i odchylenie standardowe są najczęściej używanymi miarami rozrzutu i jednymi z najważniejszych parametrów w całej statystyce.

W praktyce do obliczeń wariancji i odchylenia standardowego wykorzystujemy inne niż (3.13) i (3.14) wzory, a mianowicie zastępujemy „n" w mianowniku przez „n-1". Powodów, dla których tak czynimy jest kilka. Wyjaśnienie najważniejszego z nich wymaga krótkiej informacji o zadaniu estymacji punktowej.

Jednym z głównych celów statystyki jest wnioskowanie o całej zbiorowości statystycznej na podstawie zbadania pewnej jej części zwanej próbą. Estymacja punktowa polega na szacowaniu nieznanego parametru zbiorowości poprzez pewną wielkość, obliczoną na podstawie wyników próby. Wielkość tę nazywamy estymatorem. Przykładowo estymatorem nieznanej wariancji w całej populacji może być jej oszacowanie obliczone z próby na podstawie wzoru (3.13). Czy jednak będzie to oszacowanie optymalne? Spośród kilku wymagań stawianych estymatorom, najważniejsze są trzy: dobre estymatory powinny być efektywne, nieobciążone i zgodne.

Estymator **efektywny**, to estymator o możliwie małym rozrzucie. Jest zrozumiałe, że z populacji generalnej możemy pobierać różne próby losowe. Obliczone na podstawie poszczególnych prób wartości estymatora tego samego nieznanego parametru populacji różnią się pomiędzy sobą. Estymator będzie efektywny, jeżeli różne próby dadzą możliwie zbliżone do siebie wartości oszacowania nieznanego parametru zbiorowości.

Estymator **nieobciążony** to z kolei taki estymator, którego wartość oczekiwana jest równa wartości nieznanego parametru populacji. Innymi słowy: estymator nieobciążony szacuje nieznany parametr zbiorowości bez błędu systematycznego.

Estymator **zgodny** ma tę właściwość, że stosowanie większych liczebnie prób poprawia dokładność szacunku (estymator jest stochastycznie zbieżny do wartości szacowanego parametru).

Obydwa oszacowania wariancji: zarówno tamto z wartością n, jak i to drugie z n-1 w mianowniku, charakteryzują się takim samym stopniem efektywności i własnością zgodności. Różnica polega jedynie na tym, że estymator określony wzorem (3.13) jest lekko obciążony, zaś estymator

$$s = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}{n-1} \tag{3.15}$$

należy do grupy estymatorów nieobciążonych. Natomiast zarówno estymator (3.14) jak i estymator wyrażony wzorem

$$s = \sqrt{\frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}{n-1}} \tag{3.16}$$

są obciążonymi estymatorami odchylenia standardowego. W dalszym ciągu będziemy dla szacowania wariancji lub odchylenia standardowego z próby używać zawsze wzorów (3.15) i (3.16).

Intuicyjne wyjaśnienie stosowania wzoru (3.15) można podać wykorzystując pojęcie stopni swobody. Liczba stopni swobody jest równa liczbie niewiadomych, pomniejszonej

o liczbę niezależnych równań wiążących te niewiadome. Gdybyśmy mieli 5 niewiadomych i układ trzech równań liniowych z tymi niewiadomymi, to moglibyśmy arbitralnie przyjąć wartości dwu niewiadomych, a pozostałe trzy byłyby określone poprzez równania — układ taki miałby 2 stopnie swobody. Obliczając oszacowanie wariancji z próby o liczebności n obserwacji korzystamy „po drodze" z jednego równania, a mianowicie z zależności

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

która pozwoliła nam wyliczyć średnią. Dysponujemy wobec tego nie n lecz n-1 niezależnymi odchyleniami od średniej $\bar{x}$. Inaczej mówiąc mamy do dyspozycji n-1 stopni swobody, gdyż jeden straciliśmy dla wyznaczenia średniej — i właśnie taką wartość musimy użyć do nieobciążonego estymatora wariancji. Gdybyśmy znali *prawdziwą* średnią populacji i posługiwali się odchyleniami od tej rzeczywistej średniej — nie byłoby straty jednego stopnia swobody. Jak się przekonamy później (por. np rozdz. 7) ogólna formuła na oszacowanie wariancji da się sprowadzić do poniższego zapisu:

$$\text{oszacowanie wariancji} = \frac{\text{suma kwadratów odchyleń od pewnej wartości}}{\text{liczba stopni swobody}}$$

Na zakończenie rozważań o oszacowaniu wariancji warto podać pewną tożsamość, użyteczną przy praktycznych obliczeniach:

$$\sum (x_i - \bar{x})^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n} = \sum x_i^2 - n\bar{x}^2 \tag{3.17}$$

Użyteczność tej tożsamości wynika z faktu, że zamiast najpierw wyliczać średnią $\bar{x}$ i potem sumować odchylenia $(x_i - \bar{x})^2$ — co wymaga dwukrotnego przeglądania zbioru danych — możemy raz analizować dane, wyznaczając dwie sumy pomocnicze Σx_i oraz Σx_i^2, z których potem bez trudu określimy wartość średnią i odchylenie standardowe.

3.2.3 Obliczanie miar rozrzutu dla szeregów rozdzielczych

Zostaną teraz podane wzory na obliczanie odchylenia przeciętnego, wariancji i odchylenia standardowego dla danych w postaci szeregów rozdzielczych. Wzory te oparto na założeniu, że wszystkie obserwacje należące do danej klasy rozłożone są równomiernie na długości przedziału klasowego, a wobec tego można dla celów obliczeniowych traktować je tak, jakby były skupione w środku przedziału klasowego. Założenie takie (gdy nie jest spełnione) daje nieco zawyżone wartości miar rozrzutu.

Odchylenie przeciętne oblicza się według wzoru

$$d = \frac{\sum_{i=1}^{n} \left| \overset{\circ}{x}_i - \bar{x} \right| n_i}{\sum_{i=1}^{n} n_i} \tag{3.18}$$

gdzie:

$\overset{\circ}{x}_i$ — środek i-tego przedziału klasowego,
n — liczba klas,
n_i — liczebność i-tej klasy.

Przykład 3.1.
Obliczone dla szeregu rozdzielczego z tabeli 3.1 odchylenie przeciętne wynosi

$$d = 7{,}4 \text{ lat}$$

Oszacowanie wariancji dla szeregu rozdzielczego otrzymujemy ze wzoru

$$s^2 = \frac{\sum_{i=1}^{n} (\overset{\circ}{x}_i - \bar{x})^2 n_i}{\sum_{i=1}^{n} n_i - 1} \tag{3.19}$$

(oznaczenia jak wyżej). Od tak obliczonej wartości oszacowania wariancji w przypadku niewielkiej liczby przedziałów klasowych odejmuje się wielkość równą $h^2/12$ (gdzie: h — długość przedziału klasowego; w przypadku niejednakowych klas — długość przeciętna) zwaną poprawką Shepparda. Postępowanie takie ma na celu uniknięcie nadmiernego zawyżenia szacunku. Pierwiastkując tak poprawione oszacowanie wariancji uzyskujemy estymator odchylenia standardowego.

Przykład 3.1 (c.d.)

W naszym przypadku dla danych z tabeli 3.1 mamy
wariancja $S^2 = 76{,}66$ [rok^2]
poprawka Shepparda $h^2/12 = 10^2/12 = 8{,}33$ [rok^2]
wariancja skorygowana $s^2_{sk} = 76{,}66 - 8{,}33 = 68{,}33$ [rok^2]
odchylenie standardowe $s = 8{,}27$ lat.

Czasami charakteryzujemy wielkość rozrzutu w próbie za pomocą miernika o charakterze względnym, nazywanego współczynnikiem zmienności i zdefiniowanym jako

$$v = \frac{s}{\bar{x}} \cdot 100\% \tag{3.13}$$

Dla danych z tabeli 3.1 współczynnik ów ma wartość

$$v = (8{,}27/55{,}6) * 100\% = 14{,}9\%$$

4. ESTYMACJA PRZEDZIAŁOWA PARAMETRÓW

4.1. Ogólny problem estymacji przedziałowej

Omawiając miary rozrzutu w punkcie 3.2.2 scharakteryzowano krótko problem estymacji punktowej. Jak sobie przypominamy, chodziło o oszacowanie nieznanego parametru populacji generalnej przy pomocy pewnej pojedynczej wielkości, wyznaczonej na podstawie próby wybranej losowo z całej populacji. Jeżeli interesowała nas wariancja z populacji, to szacowaliśmy ją używając wariancji obliczonej z próby. Podobnie gdyby interesowała nas średnia z populacji, użylibyśmy jako estymatora średniej z próby. Chcąc przykładowo oszacować średnie skurczowe ciśnienie krwi u mężczyzn w wieku 30...40 lat zatrudnionych w przemyśle na stanowiskach robotniczych, można wybrać losowo próbę 31 mężczyzn, spełniających podane warunki i zmierzyć im ciśnienie krwi. Po uśrednieniu otrzyma się wartość wynoszącą np. 136 mmHg, która może być uważana za oszacowanie średniego ciśnienia skurczowego w całej rozważanej zbiorowości pracowników przemysłu. Czasami to wystarcza. Często jednak wymagane są dodatkowe informacje o dokładności szacunku. Mogą one być sformułowane na przykład w taki sposób: z dość dużym, bo wynoszącym 95% prawdopodobieństwem, przedział 136 ± 6 mmHg (tzn. od 130 do 142 mmHg) pokrywa nieznaną wartość ciśnienia skurczowego w naszej grupie pracowników przemysłu.

Estymacja przedziałowa polega właśnie na szacowaniu nieznanego parametru populacji poprzez budowanie takiego przedziału, który z zadanym z góry prawdopodobnieństwem pokrywałby nieznaną wartość parametru. Poszukiwany w zadaniu estymacji przedział nazywany jest **przedziałem ufności**, zaś ustalane a priori prawdopodobieństwo, z którym przedział ufności ma pokrywać nieznaną wartośc parametru, nosi nazwę **współczynnika (poziomu) ufności**. Współczynnik ufności podawany jest zwykle jako $1 - \alpha$ i przyjmuje najczęściej wartość 0,90, 0,95 i 0,99.

4.2. Estymacja przedziałowa średniej

Powstaje pytanie, skąd się biorą informacje pozwalające na wyznaczenie przedziału ufności. Informacji tych dostarczają dane z próby oraz teoria dotycząca rozkładu estymatorów. Stwierdzenia powyższe wyjaśnimy na przykładzie estymacji przedziałowej średniej.

Rozpoczniemy od teorii. Przede wszystkim należy sobie uzmysłowić, że średnia $\bar{x}$ obliczona z próby jest zmienną losową. Losując np. kilka 31-osobowych grup pracowników przemysłu w wieku 30...40 lat uzyskamy kilka prób pobranych z tej samej populacji. Średnie ciśnienie skurczowe $\bar{x}$ w każdej z tych grup będzie zapewne trochę inne. Uzyskane w ten sposób średnie $\bar{x}$ są kilkoma realizacjami zmiennej losowej którą, można nazwać „średnią z 31-elementowej próby pobranej losowo z populacji". Z teorii wiemy, że wartością oczekiwaną tej zmiennej losowej jest nieznana nam średnia μ populacji generalnej

$$E(\bar{x}) = \mu$$

Innymi słowy, wszystkie średnie z prób grupują się wokół rzeczywistej średniej z populacji. Grupują się tym bliżej, im mniejszy był rozrzut w samej populacji generalnej oraz im większa była liczebność próby. Z teorii wynika bowiem, że wariancja średniej z próby n-elementowej wyraża się wzorem:

$$\sigma^2(\bar{x}) = \frac{\sigma^2}{n} \tag{4.1}$$

gdzie:

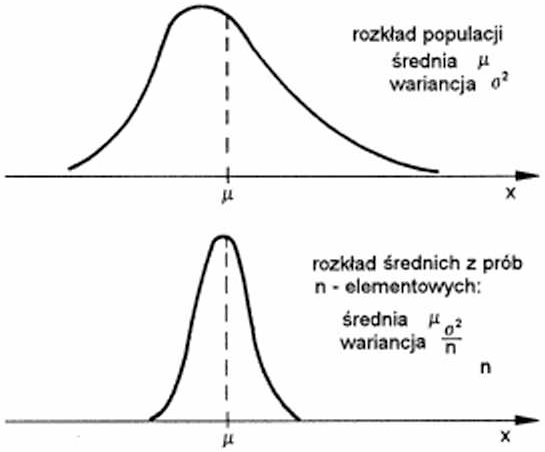
σ^2 — wariancja w populacji generalnej.

Teoria mówi nam poza tym, że jeżeli populacja ma rozkład normalny to także średnia z próby ma rozkład normalny. Co więcej, jeśli nawet rozkład populacji nie jest normalny, to rozkład średniej z próby w miarę wzrostu liczebności próby bardzo szybko dąży do rozkładu normalnego ze średnią μ (gdzie μ — rzeczywista średnia w populacji) i wariancją $\frac{\sigma^2}{n}$ (por.rys.4.1).

Wszystkie te wyniki — teoretycznie bardzo ważne dla zrozumienia własności zmiennej losowej: „średnia z próby n-elementowej" — byłyby również bardzo użyteczne dla celów estymacji przedziałowej, gdybyśmy znali rzeczywistą wartość wariancji σ^2 w populacji generalnej. Wtedy moglibyśmy skorzystać z faktu, że zmienna losowa $\bar{x}$ ma rozkład normalny ze średnią μ i wariancją $\frac{\sigma^2}{n}$, czyli że zmienna

$$u = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \tag{4.2}$$

ma znany dobrze z literatury standaryzowany rozkład normalny. Ponieważ jednak σ^2 zwykle nie jest znane, pozostaje nam wykorzystać zmienną t określoną wzorem:



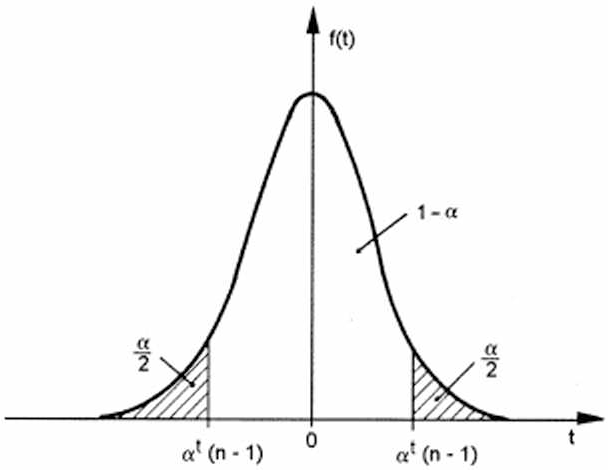
Rys. 4.1 Rozkład prawdopodobieństwa badanej cechy w populacji generalnej i rozkład średnich n-elementowych prób pobranych z tej populacji.

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \tag{4.3}$$

która w odróżnieniu od zmiennej u opiera się na dwojakich informacjach uzyskanych z próby: zawiera mianowicie (tak jak u) obliczoną z wyników próby średnią $\bar{x}$, ale ponadto jest tam oszacowanie odchylenia standardowego s obliczone także na podstawie próby, czego nie było we wzorze (4.2). Statystyka t (zmienne losowe będące funkcjami wyników próby nazywamy statystykami) nie ma już rozkładu normalnego, lecz inny dobrze zbadany rozkład, zwany rozkładem t-Studenta. Rozkłady t-Studenta to cała rodzina rozkładów scharakteryzowana przy pomocy „liczby stopni swobody". Otóż nasza zmienna t wyznaczona na podstawie próby n-elementowej posiada rozkład t-Studenta o $n-1$ stopniach swobody. Funkcja gęstości prawdopodobieństwa rozkładu t przypomina krzywą gaussowską, ale odznacza się większym rozrzutem (por. rys.4.2). Z odpowiednich tablic statystycznych wartości krytycznych rozkładu t można odczytać wartości $\alpha t_{(n-1)}$, stanowiące granice przedziału $(-\alpha t_{(n-1)}, \alpha t_{(n-1)})$, w którym to przedziale z prawdopodobieństwem $1-\alpha$ będą pojawiać się realizacje każdej zmiennej losowej o rozkładzie t z $n-1$ stopniami swobody. Nasza zmienna t (wzór (4.3)) związana z próbą pobraną z populacji także z bliskim jedności prawdopodobieństwem $1-\alpha$ znajdzie się w przedziale:

$$-\alpha t_{(n-1)} < t < \alpha t_{(n-1)}$$

Ponieważ t wyraża się zależnością (4.3), więc



Rys.4.2 Rozkład *t*-Studenta o n - 1 stopniach swobody.

$$-\alpha^{t}_{(n-1)} < \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}} < \alpha^{t}_{(n-1)}$$

Przekształcając dalej otrzymamy:

$$\frac{-\alpha^{t}_{(n-1)} \cdot s}{\sqrt{n}} < \bar{x} - \mu < \frac{\alpha^{t}_{(n-1)} \cdot s}{\sqrt{n}}$$

i ostatecznie

$$\bar{x} - \frac{-\alpha^{t}_{(n-1)} \cdot s}{\sqrt{n}} < \mu < \bar{x} + \frac{\alpha^{t}_{(n-1)} \cdot s}{\sqrt{n}} \tag{4.4}$$

Tak więc z założonym prawdopodobieństwem $1-\alpha$ spełniona jest powyższa nierówność, co oznacza, że został wyznaczony przedział ufności dla nieznanej średniej populacji. Wzór (4.4) jest słuszny dla populacji o rozkładzie normalnym i próby o niewielkiej nawet liczebości. Jeżeli liczebność próby jest duża ($n > 100$), to przedział ufności można wyznaczyć posługując się statystyką u z tą różnicą, że występującą we wzorze (4.2) wartość odchylenia standardowego populacji σ należy zastąpić oszacowaniem odchylenia standardowego s obliczonym z próby. Wówczas przedział ufności dla średniej μ jest następujący:

$$\bar{x} - \frac{\alpha^{u} \cdot s}{\sqrt{n}} < \mu < \bar{x} + \frac{\alpha^{u} \cdot s}{\sqrt{n}} \tag{4.5}$$

przy czym wartość ${}_{\alpha}u$ dla danego współczynnika ufności $1-\alpha$ wyznacza się z tablic standaryzowanego rozkładu normalnego tak by była spełniona relacja:

$$P\{-{}_{\alpha}u < u < {}_{\alpha}u\} = 1-\alpha \tag{4.6}$$

Najczęściej używane wartości ${}_{\alpha}u$ podano w tabeli 4.1

Tabela 4.1

Wartości krytyczne ${}_{\alpha}u$ rozkładu normalnego standaryzowanego

$1-\alpha$	α	${}_{\alpha}u$
0,9	0,1	1,646
0,95	0,05	1,960
0,975	0,025	2,241
0,99	0,01	2,576
0,999	0,001	3,291

Przykład 4.1.

Zmierzono ciśnienie skurczowe krwi grupie 31 mężczyzn w wieku 30...40 lat pracujących w przemyśle na stanowisku robotniczym. U jednego pracownika stwierdzono 100 mmHg, u jednego 110 mmHg, u pięciu 120 mmHg, u 7 — 130 mmHg, u 9 — 140 mmHg, u czterech — 150 mmHg, u trzech — 160 mmHg i u jednego — 170 mmHg. Znaleźć 95-procentowy przedział ufności dla ciśnienia skurczowego w rozpatrywanej populacji pracowników przemysłu. Mamy:

$n = 31$

$\bar{x} = 136{,}4 \approx 136$ mmHg

$s^2 = 243{,}7$

$s = 15{,}6$

Ponieważ próba nie jest szczególnie liczna, więc zastosujemy wzór (4.4). Otrzymamy:

$1-\alpha = 0{,}95$

$\alpha = 0{,}05$

${}_{0{,}05}t_{(30)} = 2{,}042$

$$\frac{\alpha t_{(n-1)} \cdot s}{\sqrt{n}} = 5{,}7 \approx 6 \text{ mmHg}$$

Poszukiwany 95-procentowy przedział ufności wynosi 136+6 mmHg.

Przykład 4.2.

Wyznaczyć 99-procentowy przedział ufności dla średniego wieku pacjentów chorych na raka płuc, wykorzystując dane z tabeli 3.1 i obliczenia wykorzystane w rozdziale 3. Ponieważ próba jest duża skorzystamy z metody objętej wzorem (4.5). Otrzymamy:

$n = 1357$

$\bar{x} = 55{,}6$

$s = 8{,}27$

$1 - \alpha = 0{,}99$

$\alpha = 0{,}01$

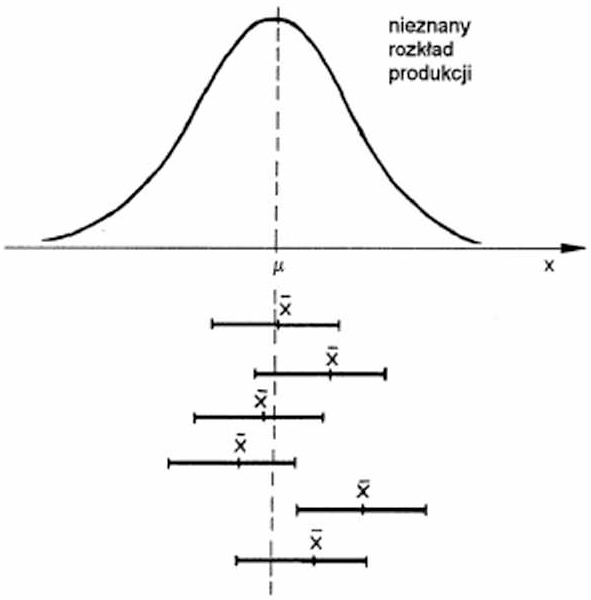
$_{0{,}01}u = 2{,}576$

$$\frac{\alpha^u \cdot s}{\sqrt{n}} = 0{,}6$$

$55 < \mu\ 56{,}2$

Wyznaczony 99-procentowy przedział ufności średniego wieku chorych jest równy 55,6 ± 0,6 lat.

Uważny czytelnik być może zauważył, że w dotychczasowym wywodzie ani razu nie użyto sformułowania typu: „z prawdopodobieństwem 0,95 nieznany parametr populacji znajdzie się w przedziale ufności". Stwierdzenie takie sugerowałoby zmienność wartości parametru populacji, podczas gdy w rzeczywistości zmienne jest usytuowanie przedziału ufności względem parametru populacji (por. np. rys. 4.3). Przykładowo: średnia całej



Rys. 4.3 Rozkład zmiennych przedziałów ufności wokół stałej wartości parametru.

populacji nie zmienia się, różne mogą być średnie obliczane z prób, różne więc może być położenie przedziału ufności otaczającego te średnie. Niekiedy zdarza się, że (z małym prawdopodobieństwem α) przedział ufności, jak to pokazano na rysunku 4.3 nie pokrywa rzeczywistej wartości estymowanego parametru.

4.3 Przedział ufności dla częstości

Załóżmy, że elementy populacji generalnej można podzielić na dwie grupy np. „*A*" i „nie *A*". Oznaczmy przez Π prawdopodobieństwo (częstość) wystąpienia elementu grupy *A* i przez $(1 - \Pi)$ prawdopodobieństwo wystąpienia elementu z grupy „nie *A*". Aby oszacować nieznaną wartość Π losujemy z populacji generalnej próbę o liczebności *n* elementów. Przez *r* oznaczymy liczbę elementów grupy *A* w próbie. Spodziewamy się, że obliczona z próby frakcja

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

będzie estymatorem nieznanej wielkości Π z populacji generalnej. I tak jest rzeczywiście. Teoria mówi nam, że wartość oczekiwana statystyki *p* jest równa Π,

$$E(p) = \Pi \tag{4.7}$$

wariancja *p* wynosi

$$\sigma^2(p) = \frac{\Pi \cdot (1 - \Pi)}{n} \tag{4.8}$$

zaś sama zmienna *p* (a właściwie zmienna $r = n \cdot p$) ma rozkład dwumianowy. W miarę jak rośnie wielkość próby *n*, rozkład *p* zmierza do normalnego.

Przy określaniu przedziału ufności dla czestości Π w populacji wykorzystujemy na ogół tę ostatnią własność, pamiętając jednak, że próba musi być duża. Wyznaczenie przedziału ufności dla częstości Π z małej próby nie jest sprawą prostą. Trzeba korzystać z własności i postaci rozkładu dwumianowego lub ze specjalnych tablic (np.: Tablice statystyczne, pod red. W. Sadowskiego). Tutaj zagadnienie to nie będzie poruszane.

Załóżmy, że rozpatrywana populacja generalna ma rozkład dwupunktowy z parametrem Π, który nie jest zbyt mały ($\Pi > 0{,}05$). Z populacji wylosowano dużą próbę ($n > 100$), przy czym wyznaczona z próby frakcja wynosi *p*. Jeżeli ani $n \cdot p$ ani $n \cdot (1 - p)$ nie są zbyt małe (są większe niż 10), to można oszacowywać przybliżony przedział ufności dla częstości Π zgodnie z poniższym zapisem:

$$P\left\{p - {}_{\alpha}u\sqrt{\frac{p(1-p)}{n}} < \Pi < p + {}_{\alpha}u\sqrt{\frac{p(1-p)}{n}}\right\} \approx 1 - \alpha \tag{4.9}$$

gdzie:

${}_{\alpha}u$ — wartość krytyczna rozkładu standaryzowanego normalnego (patrz tabela 4.1)

Przykład 4.3

Spośród studentów pewnej wyższej uczelni wylosowano do próby stu pięćdziesięciu i zapytano ich, czy palą papierosy. 105 studentów stwierdziło, że systematycznie pali papierosy. Oszacować metodą przedziałową procent palących studentów uczelni, przyjmując współczynnik ufności 0,95.

Ponieważ wszystkie podane wyżej założenia są spełnione dokonujemy obliczeń:

$$n = 150,\quad r = 105,\quad p = \frac{105}{150} = 0{,}7,\quad {}_{0{,}05}u = 1{,}96$$

$${}_{\alpha}u \cdot \sqrt{\frac{p(1-p)}{n}} = 0{,}073 \approx 0{,}07$$

$$0{,}63 < \Pi < 0{,}77$$

A zatem poszukiwany przedział ufności dla procentu palących studentów uczelni można określić jako $70 \pm 7\%$.

4.4 Przedział ufności dla wariancji

Jeżeli chcemy oszacować metodą przedziałową wariancję populacji, to powinniśmy najpierw sprawdzić, czy można uważać, że populacja ma rozkład normalny. O tym, jak to zrobić można przeczytać w rozdziale dziewiątym. Statystyka wskazuje bowiem na to, że wszystkie sposoby wnioskowania dotyczące wariancji są znacznie bardziej czułe na odchyłki rzeczywistego rozkładu od rozkładu normalnego, niż metody dotyczące średniej.

Teoria poucza nas, jak to już wiemy z punktu 3.2.2, że estymatorem nieznanej wariancji σ^2 populacji generalnej powinien być nieobciążony estymator s^2 (wzór 3.8). Wartość oczekiwana takiego estymatora jest równa wartości niezananego parametru zbiorowości:

$$E(s^2) = \sigma^2 \tag{4.10}$$

zaś wariancja s^2 dla n-elementowych prób pobranych z populacji o rozkładzie normalnym wynosi

$$\sigma^2(s^2) = \frac{\sigma^2}{n-1} \tag{4.11}$$

Badając rozkład statystyki s^2 stwierdzono, że ta zmienna losowa może być przedstawiona jako:

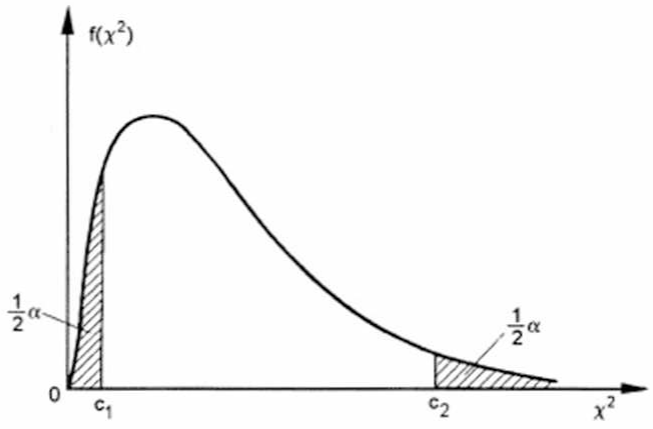
$$s^2 = \frac{\sigma^2}{n-1} * \chi^2_{(n-1)}$$

gdzie $\chi^2_{(n-1)}$ jest zmienną losową o bardzo często występującym w statystyce rozkładzie χ^2 (chi-kwadrat). Rozkład chi-kwadrat podobnie jak t-Studenta charakteryzuje się tzw. „liczbą stopni swobody". Zapis $\chi^2_{(n-1)}$ oznacza zmienną posiadającą rozkład χ^2 o $n-1$ stopniach swobody. Rozkład $\chi^2_{(n-1)}$ jest rozkładem sumy kwadratów $n-1$ standaryzowanych niezależnych zmiennych losowych o rozkładzie normalnym.

Dysponując małą, n-elementową próbą wylosowaną z populacji o rozkładzie normalnym, można przedział ufności dla wariancji σ^2 populacji generalnej określić zapisem:

$$P\left\{\frac{(n-1)s^2}{c_2} < \sigma^2 < \frac{(n-1)s^2}{c_1}\right\} = 1-\alpha \tag{4.12}$$

gdzie c_1 i c_2 są wartościami zmiennej $\chi^2_{(n-1)}$ wyznaczonymi z tablic statystycznych tak, aby spełnione były zależności (por. rys. 4.4)



Rys. 4.4 Rozkład χ^2.

$$P\left(\chi^2_{(n-1)} < c_1\right) = \frac{\alpha}{2} \quad i \quad P\left(\chi^2_{(n-1)} \geq c_2\right) = \frac{\alpha}{2}$$

Ponieważ tablice podają wartości krytyczne ${}_{\alpha}\chi^2_{(k)}$ spełniające zależność

$$P\left(\chi^2 \geq {}_{\alpha}\chi^2_{(k)}\right) = \alpha$$

więc należy odczytać z tablic wartości c_1 i c_2 jako

$$c_1 = {}_{1-\frac{\alpha}{2}}\chi^2_{(n-1)}$$

$$c_2 = {}_{\frac{\alpha}{2}}\chi^2_{(n-1)}$$

Jeżeli dysponujemy natomiast dużą próbą ($n > 100$) pobraną z populacji o rozkładzie normalnym lub zbliżonym do normalnego, to na podstawie estymatora s wyznaczonego z tej próby możemy w przybliżeniu oszacować odchylenie standardowe z populacji według poniższego wzoru:

$$P\left\{\frac{s}{1+\frac{\alpha u}{\sqrt{n}}} < \sigma < \frac{s}{1-\frac{\alpha u}{\sqrt{n}}}\right\} \approx 1-\alpha \tag{4.13}$$

gdzie ${}_{\alpha}u$ jest wartością krytyczną normalnego rozkładu standaryzowanego.

Przykład 4.4.

Na podstawie wzoru (4.13) oszacujemy metodą przedziałową odchylenie standardowe wieku pacjentów chorych na nowotwór płuc wykorzystując dane z tabeli 3.1, a także obliczenia przeprowadzone w rozdziale 3. Przyjmiemy współczynnik ufności równy 0,99. Z danych otrzymujemy:

$n = 1357$

$s = 8{,}27$

$1 - \alpha = 0{,}99$

${}_{0{,}01}u = 2{,}576$

a po wstawieniu do wzoru (4.13) uzyskamy poniższy wynik: 99- procentowe oszacowanie odchylenia standardowego wynosi

$7{,}88 < \sigma < 8{,}70$

4.5 Szacowanie niezbędnej liczebności próby

W zadaniu estymacji przedziałowej określa się szerokość przedziału ufności w oparciu o znane wyniki próby i wiadomości o rozkładzie estymatora. Często jednak po uzyskaniu n_0 obserwacji, obliczony przy założonym współczynniku ufności przedział okazuje się zbyt duży. Po to, aby uczynić szacunek dokładniejszym i nie zmieniać przyjętego wcześniej współczynnika ufności, należy zwiększyć liczebność próby. Pojawia się pytanie: ile dodatkowych obserwacji należy jeszcze uzyskać, aby szerokość przedziału ufności była nie większa od założonej z góry wielkości.

Odpowiedzi na takie pytania są szczególnie proste, gdy rozpiętość przedziału ufności można łatwo przedstawić jako funkcję liczebności próby. Przykładowo przy odpowiednich założeniach przedziały ufności dla średniej i częstości można zapisać jako:

$$\bar{x} \pm d \qquad \text{lub} \qquad \Pi \pm d$$

gdzie:

$$d = {}_{\alpha}t\frac{s}{\sqrt{n}} \qquad \text{oraz} \qquad d = \alpha u\sqrt{\frac{p(1-p)}{n}}$$

Wyliczając z tych wzorów n otrzymuje się:

$$n = \frac{\alpha t^2 \cdot s^2}{d^2} \qquad oraz \qquad n = \frac{\alpha u^2 \cdot p(1-p)}{d^2} \tag{4.14}$$

Po uzyskaniu pierwszych n_0 obserwacji, wykorzystując wzory (4.14) można obliczyć liczbę wszystkich obserwacji n niezbędnych dla osiągnięcia przedziału ufności o założonej z góry rozpiętości wynoszącej $2d$. We wzorach (4.14) należy wykorzystać oszacowania s i p obliczone z posiadanych n_0 obserwacji, zaś ${}_{\alpha}t$ odczytać z tablic dla $n_0 - 1$ stopni swobody. Jeżeli ustali się pewne $n > n_0$, to należy dodatkowo uzyskać $(n - n_0)$ obserwacji.

Przykład 4.5. [Parker]

W badaniu taksonomicznym chcemy oszacować długość określonej struktury z danej populacji z dokładnością ±1.0 mm przy poziomie ufności 0,95. Po wykonaniu 26 pomiarów oszacowano odchylenie standardowe, które wyniosło 4 mm. Ile pomiarów należy jeszcze wykonać?

Mamy

$$d = 1 \qquad s = 4 \qquad {}_{0,05}t_{(25)} = 2{,}06 \qquad n_0 = 26$$

$$n = \frac{(2{,}06)^2 \cdot 4^2}{1^2} = 69{,}32 \approx 70$$

Należy jeszcze wykonać dodatkowo $n - n_0 = 70 - 26 = 44$ pomiary.

Przykład 4.6

Kontynuując postępowanie z przykładu 4.3 określić z iloma studentami należy dodatkowo przeprowadzić wywiad, aby oszacować procent palących z dokładnością co najwyżej ±5% przy współczynniku ufności $1 - \alpha = 0{,}95$. Mamy

$$n_0 = 150 \qquad p = 0{,}7 \qquad {}_{0{,}05}u = 1{,}96 \qquad d = 0{,}05$$

$$n = \frac{(1{,}96)^2 \cdot 0{,}7 \cdot 0{,}3}{(0{,}05)^2} = 322{,}7 = 323$$

A zatem należy jeszcze dodatkowo przeprowadzić wywiad na temat palenia z $n - n_0 = 323 - 150 = 173$ studentami.

5. PARAMETRYCZNE TESTY ISTOTNOŚCI

5.1 Testowanie hipotez statystycznych

Zasadniczą domeną statystyki jest weryfikacja hipotez statystycznych, czyli pewnych przypuszczeń dotyczących rozkładu populacji testowanych przy wykorzystaniu wyników próby losowej pobranej z tej populacji. Zwykło się dzielić hipotezy statystyczne na dwie grupy: hipotezy **parametryczne** i hipotezy **nieparametryczne**. Hipotezy pierwszej grupy związane są z wartościami parametrów rozkładów populacji. Im to właśnie poświęcona będzie większa część niniejszego podręcznika. Hipotezy nieparametryczne dotyczą generalnie typu rozkładu populacji. Również one są często wykorzystywane w badaniach biologicznych i medycznych, będą więc omówione w jednym z dalszych rozdziałów.

Proces weryfikacji hipotezy statystycznej przebiega według pewnego wzorca postępowania zwanego **testem statystycznym**. Gdy weryfikacji podlega hipoteza parametryczna mówimy o teście parametrycznym. Test rozpoczyna się od postawienia tej hipotezy, która będzie podlegała sprawdzaniu — hipoteza taka nosi nazwę hipotezy zerowej i bywa oznaczana H_0. Następnie musi zostać sformułowana hipoteza alternatywna H_1 — konkurencyjna względem hipotezy zerowej. Jeżeli bowiem w trakcie testowania hipoteza zerowa zostanie odrzucona jako nieprawdziwa, to będzie przyjęta hipoteza alternatywna. Sposób konstrukcji konkretnego postępowania przy testowaniu zależy od istoty badanego problemu statystycznego, jednakże zawsze weryfikacja przebiega tak, aby zapewnić możliwie małe prawdopodobieństwo pomyłek. Możliwe do popełnienia błędy dzielimy na błędy pierwszego rodzaju, polegające na odrzuceniu hipotez w gruncie rzeczy prawdziwych oraz na błędy drugiego rodzaju, spowodowane przyjęciem hipotez fałszywych.

Decyzja o przyjęciu bądź odrzuceniu hipotezy statystycznej nie jest tożsama z logicznym udowodnieniem jej prawdy lub fałszu. Dlaczego bowiem odrzucamy hipotezę? Dlatego, że analiza wyniku próby wskazuje na to, że przy takiej próbie jest bardzo mało prawdopodobne, aby hipoteza dotycząca populacji była prawdziwa. Rzeczywistość reprezentowana przez próbę nie pasuje, nie zgadza się z teorią, której reprezentantem jest hipoteza zerowa, więc tę ostatnią odrzucamy. Czyniąc tak liczymy się jednak z tym, że bardzo rzadko, ale jednak czasami hipoteza zerowa rzeczywiście jest słuszna, a tylko wynik próby jest przypadkowo taki właśnie, jaki jest, tzn. mało prawdopodobny przy tej hipotezie.

Największe znaczenie w statystyce posiadają tzw. **testy istotności**. W wyniku ich przeprowadzenia możliwa jest tylko decyzja o odrzuceniu hipotezy zerowej i przyjęciu hipotezy alternatywnej. W przeciwnym przypadku nie ma możliwości podjęcia decyzji o przyjęciu hipotezy zerowej, a jedynie formułuje się stwierdzenie, że brak jest podstaw do odrzucenia tej hipotezy. Jest tu pewna analogia do sytuacji sędziego: hipoteza zerowa to założenie niewinności oskarżonego. Sędzia może uznać, że wina została sprawcy udowodniona i wydać wyrok skazujący. Będzie to odrzucenie hipotezy zerowej i przyjęcie hipotezy alternatywnej. Jeżeli jednak sędzia stwierdzi, że wina nie została udowodniona, to nie będzie to równoznaczne ze stwierdzeniem niewinności oskarżonego, czyli przyjęciu hipotezy zerowej. Tylko takie możliwości ostatecznego werdyktu uwarunkowane są tym, że w procesie testowania brane jest pod uwagę tylko prawdopodobieństwo popełnienia błędu pierwszego rodzaju, który jak wiemy polega na odrzuceniu hipotezy prawdziwej. To założone z góry przed rozpoczęciem procesu weryfikacji hipotezy małe prawdopodobieństwo błędu nazywane jest poziomem istotności i oznaczane literą α. Najczęściej przyjmowane wartości poziomu istotności, to 0,05 oraz 0,1 i 0,01.

Taki wydawałoby się niepełny logicznie sposób rozstrzygania problemów hipotez statystycznych właściwie nie jest wielkim ograniczeniem. Na ogół hipotezy zerowe nie są tymi, na których badaczowi szczególnie zależy. Formułowane są one zwykle w sposób „neutralny", np. że nie ma różnic między dwiema próbami, że populacje, z których próby zostały pobrane są identyczne, że nie ma związku między dwoma badanymi zjawiskami. Są to wszystko mało użyteczne, mało konstruktywne i zupełnie nietwórcze sformułowania dla badacza — eksperymentatora. On właśnie pragnie wykazać istnienie różnic, związków i zależności wzajemnych, czyli jemu właściwie zależy na udowodnieniu hipotezy odwrotnej niż ta, którą postawiono, czyli na obaleniu hipotezy pierwotnej. Takie właśnie przewrotne narzędzie jest w stanie zaoferować badaczowi statystyka. Dzięki jego wykorzystaniu eksperymentator uzyskuje silne potwierdzenie swoich wyników, jeżeli tylko znajdą się statystyczne podstawy do odrzucenia niechcianej hipotezy zerowej.

Jaka jest zasadnicza idea samego procesu wnioskowania statystycznego podczas testowania hipotez? Otóż w zależności od tego, co właściwie ma być badane, czyli od sformułowania hipotezy H_0 tworzy się pewną statystykę Z opartą na wynikach próby złożonej z n obserwacji. Następnie posługując się teorią oraz robiąc bardzo ważne założenia, że mianowicie hipoteza H_0 dotycząca populacji z której wylosowano próbę jest prawdziwa — wyznacza się rozkład statystyki Z. Dalej określa się jakie wartości musiałaby przyjmować zmienna losowa Z, aby było to mało prawdopodobne. „Mało prawdopodobne" w poprzednim zdaniu oznacza, że prawdopodobieństwo zaistnienia tych wartości statystyki Z byłoby równe ustalonej z góry małej wielkości α, zwanej jak wiemy poziomem istotności. Te mało prawdopodobne wartości statystyki Z tworzą tzw. obszar krytyczny Q spełniający zależność

$$P\{Z \subset Q\} = \alpha$$

Tym obszarem krytycznym może być przedział lub np. zbiór składający się z dwóch rozłącznych przedziałów. Dalej rozumuje się tak: jeżeli obliczona z próby wartość statystyki Z znalazła się w obszarze krytycznym, to nastąpiło zdarzenie bardzo mało prawdopodobne. Zdarzenie takie właściwie nie powinno zaistnieć. Skoro jednak zaszło, to musiały zostać poczynione błędy w rozumowaniu prowadzącym do określenia prawdopodobieństwa trafienia przez zmienną Z do obszaru krytycznego. Ponieważ do teorii mamy zaufanie, więc widocznie nie jest spełnione założenie o prawdziwości hipotezy H_0, które to założenie posłużyło do określenia obszaru krytycznego. Hipotezę H_0 odrzucamy więc przyjmując hipotezę alternatywną H_1. Oczywiście ponieważ z małym prawdopodobieństwem α założenie o prawdziwości H_0 jest spełnione (mimo że Z wpada do obszaru krytycznego), więc odrzucając H_0 popełniamy niekiedy błąd pierwszego rodzaju. Jeżeli jednak obliczona z próby wartość statystyki Z znalazła się poza obszarem krytycznym, to przy założeniu prawdziwości H_0, zaszło zdarzenie o dużym prawdopodobieństwie $1-\alpha$, więc nie ma żadnych podstaw do kwestionowania założenia i odrzucania H_0.

W następnych punktach omówione zostaną najczęściej stosowane paramertryczne testy istotności. Nieco bardziej szczegółowo opiszemy test istotności dla średniej, aby opis ten mógł stanowić przykład przedstawionej ogólnej ideologii weryfikacji hipotez.

5.2. Test istotności dla średniej.

Z populacji o rozkładzie normalnym wylosowano n-elementową próbę (może to być próba mała, tzn. taka że $n < 30$). Na podstawie znajomości wyników tej próby należy na poziomie istotności α zweryfikować hipotezę zerową mówiącą, że średnia populacji μ jest równa pewnej ustalonej wartości μ_0

$$H_0 : \mu = \mu_0 \tag{5.1}$$

wobec hipotezy alternatywnej

$$H_1 : \mu \neq \mu_0 \tag{5.2}$$

Znajomość wyniku próby pozwala wyznaczyć średnią $\bar{x}$ oraz oszacowanie odchylenia standardowego s. Spodziewamy się, że statystyka zaangażowana w procesie testowania rozpatrywanej hipotezy będzie ściśle związana ze średnią z próby $\bar{x}$. Jak pamiętamy z punktu 4.2 średnia $\bar{x}$ jest zmienną losową o wartości oczekiwanej μ, odchyleniu standardowym $\sigma/\sqrt{n}$, (gdzie σ — odchylenie standardowe całej populacji) i rozkładzie normalnym. Wobec tego statystyka u

$$u = \frac{\bar{x} - \mu}{\sigma} \sqrt{n} \tag{5.3}$$

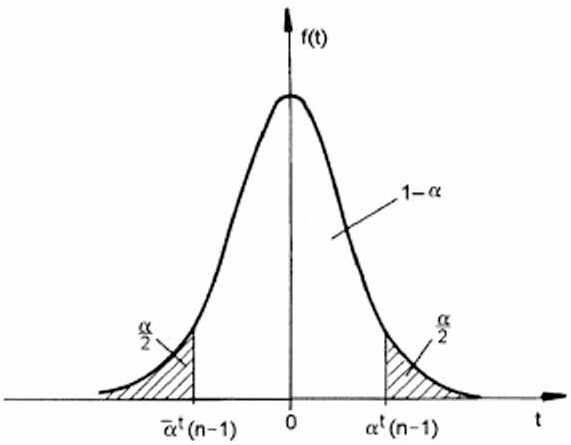
ma dobrze znany rozkład normalny standaryzowany. Nie możemy z tego faktu bezpośrednio skorzystać, gdyż nie znamy σ. Tworzymy więc statystykę t

$$t = \frac{\bar{x} - \mu}{s} \sqrt{n} \tag{5.4}$$

o której wiemy, że posiada rozkład t-Studenta o $n - 1$ stopniach swobody. Teraz przyjmujemy założenia o prawdziwości hipotezy zerowej,tzn. traktujemy że średnią populacji jest wartość μ_0. Wykorzystana w procesie weryfikacji hipotezy statystyka t będzie więc po przyjęciu tego założenia określana jako

$$t = \frac{\bar{x} - \mu_0}{s} \sqrt{n} \tag{5.5}$$

Rysunek 5.1 pokazuje rozkład zmiennej t. Zaznaczono na nim obszar krytyczny testu. Obszar ten stanowią dwa rozłączne przedziały dające się opisać zależnością



Rys. 5.1. Dwustronny obszar krytyczny w tekście średniej wykorzystującym statystykę t.

$$|t| \geq \alpha t_{(n-1)} \tag{5.6}$$

i posiadające tę właściwość, że prawdopodobieństwo osiągnięcia przez zmienną t wartości z tego obszaru jest małe i wynosi α

$$P\left(|t| \geq \alpha t_{(n-1)}\right) = \alpha$$

Wartość ${}_{\alpha}t_{(n-1)}$ odczytujemy z tablic. Następnie obliczamy wartość statystyki t (wzór(5.5)) wykorzystując wyznaczone uprzednio z próby wielkości $\bar{x}$ i s. Gdyby hipoteza zerowa była prawdziwa, to obliczona wartość t z dużym prawdopodobieństwem $1-\alpha$ znalazłaby się poza obszarem krytycznym. A więc jeżeli rzeczywiście t wypadnie poza obszarem krytycznym

$$|t| < {}_{\alpha}t_{(n-1)} \tag{5.7}$$

to nie ma podstaw do odrzucenia hipotezy H_0. Jeżeli natomiast t obliczone z próby znajduje się w obszarze krytycznym (5.6), to mamy do czynienia z bardzo mało prawdopodobnym zdarzeniem. Zdarzenie to ma tak małe prawdopodobieństwo (α), że właściwie nie powinno nastąpić. Skoro jednak zaistniało, to należy powątpiewać w słuszność założeń przyjętych w trakcie całego rozumowania mającego na względzie ustalenie parametrów rozkładu statystyki t. Jedynym założeniem dającym się obalić jest to o prawdziwości hipotezy $H_0 : \mu = \mu_0$. Odrzucamy więc hipotezę zerową i przyjmujemy alternatywną H_1, mówiącą że średnia populacji jest różna od liczby μ_0. Oczywiście czyniąc tak możemy czasami popełnić błąd, gdyż wówczas gdy hipoteza H_0 jest słuszna, to z małym prawdopodobieństwem α wartość t może znaleźć się jednak w obszarze krytycznym.

Postać obszaru krytycznego testu zależy od hipotezy alternatywnej. Powyższy test stosujemy wówczas, gdy hipoteza alternatywna ma postać $\mu \neq \mu_0$. Test taki nazywamy dwustronnym, a obszar krytyczny składa się wtedy z dwóch rozłącznych przedziałów. Możemy również budować testy jednostronne. Gdyby hipoteza alternatywna miała postać

$$H_1 : \mu < \mu_0$$

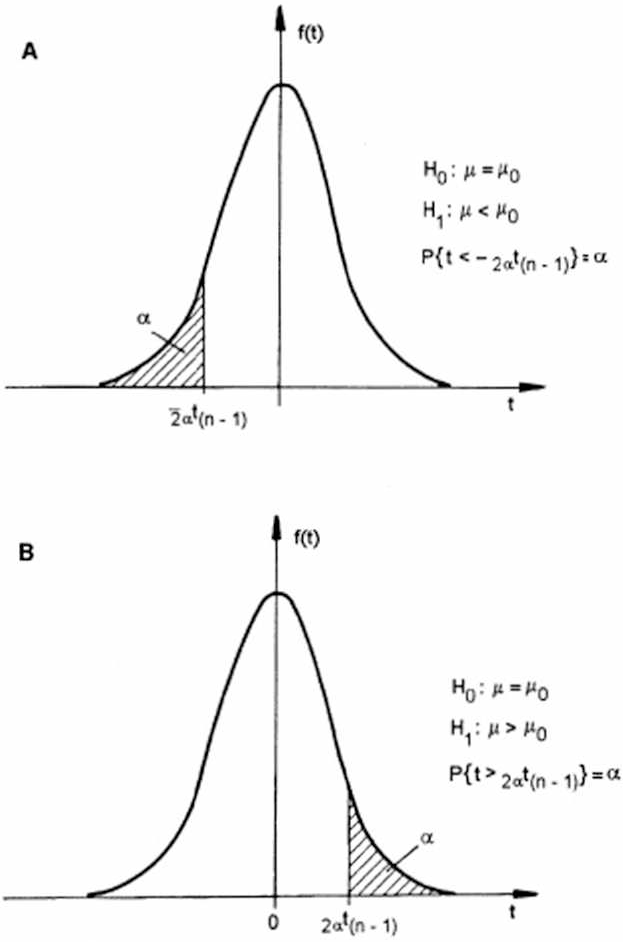
to stosowalibyśmy test lewostronny, dla którego obszar krytyczny określalibyśmy tak, aby było spełnione (por. rys.5.2):

$$P\{t \leq t(\alpha)\} = \alpha$$

Będzie to pojedynczy przedział na osi liczbowej. Ponieważ tablice statystyczne są skonstruowane dla testów dwustronnych, więc jako górną granicę obszaru krytycznego należy przyjąć odczytaną z tablic testu dwustronnego wartość ${}_{2\alpha}t_{(n-1)}$ ze znakiem ujemnym

$$t(\alpha) = -{}_{2\alpha}t_{(n-1)}$$

W teście lewostronnym odrzucamy hipotezę zerową, gdy t obliczone z próby spełnia zależność



Rys. 5.2 Obszary krytyczne dla jednostronnych testów średniej.
A — obszar lewostronny,
B — obszar prawostronny.

$$t \leq -_{2\alpha} t_{(n-1)}$$

Analogicznie dla hipotezy alternatywnej

$$H_1 : \mu > \mu_0$$

budujemy prawostronny obszar krytyczny

$$P\{t \geq t(\alpha)\} = \alpha$$

gdzie $t(\alpha) = {}_{2\alpha}t_{(n-1)}$ i odrzucamy hipotezę zerową, gdy $t \geq {}_{2\alpha}t_{(n-1)}$

Przykład. 5.1.

Załóżmy że nasza próba zawiera 21 obserwacji. Aby odrzucić hipotezę zerową w teście dwustronnym przy $\alpha = 0{,}05$ trzeba, aby t spełniało

$$|t| \geq {}_{0,05}t_{(20)} = 2{,}086$$

Przy tym samym poziomie istotności i teście jednostronnym na to aby odrzucić H_0 wystarcza by

$$t \leq -{}_{0,1}t_{(20)} = -1{,}725 \qquad \text{lub} \qquad t \geq {}_{0,1}t_{(20)} = 1{,}725 \ .$$

Statystyka t zależy liniowo od różnicy średniej z próby i hipotetycznej średniej populacji, a więc jak widać z powyższych wartości granic obszaru krytycznego łatwiej jest odrzucić hipotezę zerową w teście jednostronnym. Nie należy jednak zbyt łatwo ulegać pokusie. Trzeba pamiętać, że decyzja o stosowaniu testu jednostronnego nie może być podejmowana dopiero po zapoznaniu się z danymi i zaobserwowaniu kierunku odchyleń. Test jednostronny możemy stosować, gdy przeprowadzana jeszcze przed uzyskaniem próby analiza badanego zjawiska utwierdziła nas w przekonaniu, że np. odchylenia w kierunku wartości mniejszych są zawsze pochodzenia losowego, niezależnie od tego, jakie są duże, zaś odchylenia w kierunku przeciwnym poza składową losową mogą być powodowane działaniem jakiegoś dodatkowego czynnika. W takim przypadku uzasadnione jest przeprowadzenie testu prawostronnego dla potwierdzenia istotności (czasami mówimy znamienności) wpływu tego czynnika na obserwowane wartości. Ponieważ w praktyce z taką sytuacją mamy rzadko do czynienia, zapamiętajmy, że prawie zawsze należy stosować dwustronne testy istotności.

I jeszcze jedna uwaga o charakterze praktycznym. **Im mniejszy poziom istotności, tym trudniej odrzucić hipotezę H_0.** Stąd na ogół dopuszcza się pewien kompromis pomiędzy możliwością pomyłki a możliwością efektywnego wykorzystania aparatu statystycznego. W wyniku tego kompromisu na ogół stosuje się poziom istotności $\alpha = 0{,}05$. W wyjątkowych wypadkach, przy bardzo ważnych badaniach medycznych, testowanie przeprowadza się na niższym poziomie istotności, np. $\alpha = 0{,}01$.

Jeśli populacja generalna ma rozkład normalny lub zbliżony do normalnego, a liczebność próby jest znaczna ($n > 100$) to hipotezę $H_0 : \mu = \mu_0$ wobec hipotezy alternatywnej np. $H_1 : \mu \neq \mu_0$ można weryfikować stosując statystykę u

$$u = \frac{\overline{x} - \mu_0}{s}\sqrt{n}$$

charakteryzującą się w przybliżeniu rozkładem normalnym standaryzowanym. Dalsze postępowanie jest analogiczne, jak w przypadku testu poprzednio opisanego, ${}_{\alpha}u$ można odczytać z tabeli wartości krytycznych rozkładu normalnego standaryzowanego (tab. 4.1).

Przykład 5.2 (wg [Parker])

W doświadczeniu biochemicznym bada się czas życia żywych komórek w toksycznym środowisku. Rozkład tego czasu można przyjąć za normalny. Dokonano ośmiu pomiarów i otrzymano następujące czasy życia komórek w badanym środowisku (w godzinach): 4,7; 5,3; 4,0; 3,8; 6,2; 5,5; 4,5; 6,0. Przyjmując poziom istotności $\alpha = 0{,}05$ sprawdzić hipotezę, że średni czas życia tych komórek w tym środowisku wynosi 4,0 godziny. Otrzymujemy:

$\mu_0 = 4{,}0$
$\bar{x} = 5{,}0$
$n = 8$
$s^2 = 0{,}794$
$s/\sqrt{n} = 0{,}315$
$t = 3{,}17$
${}_{0{,}05}t_{(7)} = 2{,}365$

Ponieważ $|t| \geq {}_{0{,}05}t_{(7)}$ hipotezę $H_0 : \mu = 4{,}0$ należy odrzucić przyjmując, że wartość średnia czasu życia komórek jest istotnie różna od 4 godzin.

Zauważymy, że ten sam wynik otrzymalibyśmy obliczając przedział ufności dla μ przy współczynniku ufności $1 - \alpha = 0{,}95$. Otrzymamy wówczas przedział

$$5{,}0 \pm 0{,}75$$

który nie obejmuje hipotetycznej wartości średniego czasu życia komórek równej 4 godziny.

5.3. Testowanie różnicy między dwiema średnimi

Badania wartości średnich w dwóch populacjach są jednymi z najczęściej wykonywanych testów statystycznych w biologii i medycynie. Przykładem mogą być porównania efektów nowej metody leczenia ze starą, czy wybranych cech biochemicznych populacji ludzi zdrowych i chorych.

Załóżmy, że rozpatrujemy dwie populacje generalne o rozkładach normalnych. Nie są znane wartości parametrów rozkładu tych populacji, jedynie można przyjąć, że wariancje są w obu populacjach takie same. Z obu populacji wylosowano niezbyt duże próby o liczebnościach odpowiednio n_1 i n_2. Należy zweryfikować hipotezę zerową mówiącą, że średnie w obu populacjach są równe

$$H_0 : \mu_1 = \mu_2$$

wobec hipotezy alternatywnej postulującej nierówność średnich

$$H_1 : \mu_1 \neq \mu_2$$

Do testowania wykorzystujemy statystykę t

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \tag{5.9}$$

gdzie:

$\bar{x}_1$ i $\bar{x}_2$ — średnie odpowiednio z pierwszej i drugiej próby,

s^2 — oszacowanie wspólnej wartości wariancji σ^2 obu populacji uzyskane na podstawie estymatorów wariancji s_1^2 i s_2^2 z obu prób, wyrażające się wzorem

$$s_2 = \frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{(n_1 - 1) + (n_2 + 1)} \tag{5.10}$$

Statystyka (5.9) ma rozkład t-Studenta o $n_1 + n_2 - 2$ stopniach swobody. Konstrukcja obszaru krytycznego i dalsze postępowanie jest analogiczne jak w teście omówionym w punkcie 5.2. W tym i dalszych przypadkach będziemy przedstawiać opisy jedynie wariantu dwustronnego stosownych testów. Każdorazowo jednak, gdy tylko znajduje to odpowiednią motywację, można używać testów jednostronnych modyfikując odpowiednio hipotezę alternatywną i dalszy tok postępowania, tak jak to przedstwiono w punkcie 5.2.

Czasem mamy do czynienia z dwoma próbami, które można traktować jako zbiory obserwacji dotyczących tych samych obiektów. Przykładowo niech x_i oraz y_i będą wartościami pewnej cechy oznaczanej u n pacjentów odpowiednio przed i po kuracji. Wówczas zamiast liczyć średnie osobno z wyników przed oraz po kuracji i testować hipotezę $H_0 : \mu_x = \mu_y$ lepiej jest najpierw dla każdego pacjenta (obiektu) obliczyć różnicę

$$z_i = x_i - y_i \tag{5.11}$$

a następnie weryfikować hipotezę o tym, że wartość średnia różnic jest w całej populacji równa zeru

$$H_0 : \mu_z = 0$$

Jest to tak zwany test dla par danych. Wykorzystujemy w nim statystykę t

$$t = \frac{\bar{z}}{s_z}\sqrt{n} \tag{5.12}$$

gdzie:

$\bar{z}$ — wartość średnia różnic z_i,

s_z — oszacowanie odchylenia standardowego różnic według znanego wzoru

$$s_z = \sqrt{\frac{\sum (z_i - \bar{z})^2}{n-1}} \tag{5.13}$$

Statystyka (5.12) ma rozkład t o $n-1$ stopniach swobody. Dalsze postępowanie znamy. Korzyścią ze stosowania tego testu jest wyeliminowanie wpływu różnic między poszczególnymi obiektami (np różnic osobniczych między pacjentami), dlatego też wynik testu można uważać za bardziej obiektywny.

Jeżeli tylko są możliwości, należy stosować test dla par, a nie poprzednio omówiony test porównywania dwóch średnich. Jeżeli jednak musimy porównywać dwie średnie, a dodatkowym utrudnieniem jest niemożność przyjęcia założenia o równości wariancji w obu populacjach (por. punkt 5.7), z których wylosowano próby, to gdy obie próby mają dużą liczebność (n_1, $n_2 > 100$) można wykorzystywać statystykę u

$$u = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{5.14}$$

(oznaczenia jak we wzorach (5.9) i (5.10)) mającą w przybliżeniu rozkład normalny standaryzowany. Jeżeli zaś dysponujemy małymi próbami lub próby znacznie różnią się liczebnością, to lepiej stosować statystykę t:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{5.15}$$

która ma w przybliżeniu rozkład t-Studenta o liczbie stopni swobody ν danej wzorem

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}{\left(\frac{s_1^2}{n_1}\right)^2 \frac{1}{n_1+1} + \left(\frac{s_2^2}{n_2}\right)^2 \frac{1}{n_2+1}} - 2 \tag{5.16}$$

Wyliczoną ze wzoru (5.16) wartość należy zaokrąglić do najbliższej liczby naturalnej.

Przykład 5.3.

W próbie klinicznej nowego środka do leczenia anurezy każdy z 29 pacjentów przez 14 dni otrzymywał lek, a przez inne 14 — placebo. Kolejność przyjmowania tych środków była dla każdego pacjenta ustalana losowo. Tabela 5.1 przedstawia liczbę „suchych" nocy stwierdzoną w ciągu każdego okresu leczenia Porównanie skuteczności leku i placebo przeprowadzimy testem dla danych sparowanych.

Tabela 5.1 (według [Armitage])

Liczba „suchych" nocy (na 14 badanych) u pacjentów otrzymujących lek i placebo

L.p.	(1) lek	(2) placedo	(3) różnica (1) – (2)	L.p.	(4) placedo	(5) lek	(6) różnica (5) – (4)
1	8	5	3	2	12	11	–1
3	14	10	4	5	6	8	2
4	8	0	8	8	13	9	–4
6	9	7	2	10	8	8	0
7	11	6	5	12	8	9	1
9	3	5	–2	14	4	8	4
11	6	0	6	15	8	14	6
13	0	0	0	17	2	4	2
16	13	12	1	20	8	13	5
18	10	2	8	23	9	7	–2
19	7	5	2	26	7	10	3
21	13	13	0	29	7	6	–1
22	8	10	–2				
24	7	7	0				
25	9	0	9				
27	10	6	4				
28	2	2	0				

Mamy:

$n = 29$

$\bar{z} = 2{,}172$

$s_z^2 = 11{,}005$

$t = 3{,}53$

$_{0,05}t_{(28)} = 2{,}048$

$|t| > {}_{0,05}t_{(28)}$

Różnica między skutecznością leku i placebo jest wysoce istotna.

5.4 Test istotności dla częstości

Zakładamy, że populacja generalna ma rozkład dwupunktowy z nieznanym prawdopodobieństwem (częstością) Π występowania elementów z grupy A. Z populacji wylosowano dużą próbę n elementów ($n > 100$), przy czym okazało się, że w próbie tej jest r elementów grupy A. W oparciu o wyniki tej próby należy zweryfikować hipotezę zerową, że wartość nieznanej częstości Π w populacji wynosi Π_0 ($H_0 : \Pi = \Pi_0$) wobec hipotezy alternatywnej $H_1 : \Pi \neq \Pi_0$).

Frakcję elementów grupy A w próbie oznaczymy przez p:

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

Do weryfikacji hipotezy H_0 wykorzystamy statystykę u

$$u = \frac{p - \Pi}{\sqrt{\frac{\Pi_0 (1 - \Pi_0)}{n}}} \tag{5.17}$$

która ma w przybliżeniu rozkład normalny standaryzowany. Dalsze postępowanie jest analogiczne jak w poprzednich testach.

5.5. Porównywanie dwóch częstości

Większe znaczenie praktyczne od poprzedniego testu ma kolejny, który umożliwia porównywanie dwu częstości. Badaniu podlegają dwie populacje generalne o rozkładach dwupunktowych, przy czym Π_1 oznacza prawdopodobieństwo (częstość) występowania elementów grupy A w pierwszej populacji, zaś Π_2 — w drugiej. Z każdej populacji wylosowano niezależnie dużą próbę. W próbie o liczebności n_1 pochodzącej z populacji pierwszej stwierdzono r_1 elementów grupy A. Próba o liczebności n_2 pobrana z drugiej populacji zawierała r_2 elementów grupy A. Na podstawie takich wyników należy sprawdzić hipotezę zerową o równości częstości w obu populacjach ($H_0 : \Pi_1 = \Pi_2$) wobec hipotezy alternatywnej $H_1 : \Pi_1 \neq \Pi_2$. Oznaczymy frakcje z prób przez p_1 i p_2

$$p_1 = \frac{r_1}{n_1} \qquad p_2 = \frac{r_2}{n_2}$$

Obliczamy wartość uśrednioną frakcji z „połączonej" próby

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

oraz „pseudoliczebność" „połączonej" próby

$$n = \frac{n_1 \cdot n_2}{n_1 + n_2}$$

Następnie formułujemy statystykę u

$$u = \frac{p_1 - p_2}{\sqrt{\frac{p\,(1-p)}{n}}} \tag{5.18}$$

która w przybliżeniu charakteryzuje się rozkładem normalnym standaryzowanym. Statystyka (5.18) może być w znany nam już z poprzednich podrozdziałów sposób wykorzystywana do weryfikacji hipotezy o równości dwóch częstości.

Czasami dysponujemy obserwacjami tworzącymi n par. Każda para obserwacji związana jest z tym samym obiektem. Para obserwacji to wynik dwu prób, z których każda mogła dać rezultat „A" lub „nie A". A oto zestawienie możliwych wyników:

Wyniki		Liczba par
Próba 1	Próba 2	
A	A	k
A	nie A	v
nie A	A	w
nie A	nie A	m
		Razem: n

Frakcje elementów A w próbie 1 i w próbie 2 wynoszą odpowiednio:

$$p_1 = \frac{k+v}{n} \quad \text{i} \quad p_2 = \frac{k+w}{n}$$

zaś ich różnica

$$p_1 - p_2 = \frac{v - w}{n}$$

Testujemy hipotezę zerową mówiącą o tym, że różnica między częstościami rezultatów typu A w populacjach związanych z obu próbami jest równa zeru. Wykorzystywana do weryfikacji tej hipotezy statystyka u

$$u = \frac{v - \frac{1}{2}(v + w)}{\frac{1}{2}\sqrt{v + w}} \tag{5.19}$$

ma w przybliżeniu rozkład normalny standaryzowany. Jak widać test przeprowadzony z wykorzystaniem tej statystyki opiera się tylko na informacjach związanych z parami, w których występują różne rezultaty obu prób. Test ten (zwany testem McNemara) należy stosować we wszystkich tych przypadkach, gdzie jest to dopuszczalne ze względu na rodzaj danych.

Statystyki (5.19) — w odróżnieniu od wszystkich statystyk przedstawianych w poprzednich testach istotności — nie można użyć do określenia przedziału ufności dla różnicy częstości $\Pi_1 - \Pi_2$. Przedział taki można zbudować stosując formułę

$$\frac{v - w}{n} \pm \alpha u \frac{\sqrt{v + w}}{n} \tag{5.20}$$

Przykład 5.4 (według [Armitage])

Sto próbek plwociny posiano na dwóch różnych podłożach A i B. Zadanie polega na porównaniu zdolności tych dwu podłoży do wykrywania prątków gruźlicy. Wyniki przedstawiono w tabeli 5.2. Testujemy hipotezę zerową, że przydatność obu podłoży jest taka sama. Stosujemy test dwustronny McNemara. Mamy:

$v = 24$

$w = 10$

$n = 100$

$$u = \frac{24 - \frac{1}{2}(24 + 10)}{\frac{1}{2}\sqrt{24 + 10}} = 2{,}401$$

$_{0{,}05}u = 1{,}960$

Tabela 5.2

Rozkład 100 próbek plwociny w zależności od posiewów na dwu różnych podłożach

		Podłoże B		
		+	–	Razem
Podłoże A	+	40	24	64
	–	10	26	36
	Razem	50	50	100

Ponieważ $|u| > {}_{0,05}u$, więc hipotezę o jednakowej przydatności obu podłoży należy odrzucić. Oszacowany według (5.20) 95-procentowy przedział ufności dla różnicy częstości wynosi

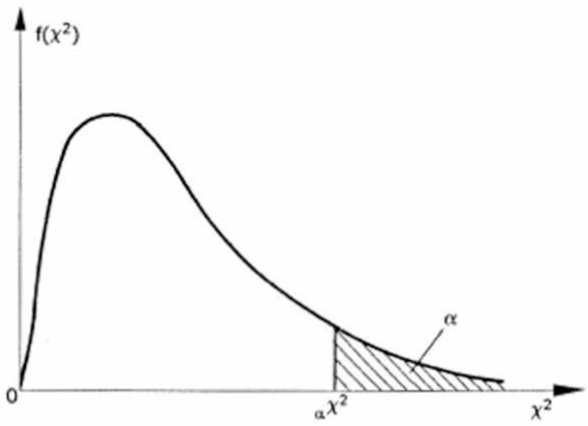
$$\frac{24-10}{100} \pm \frac{1{,}96\sqrt{24+10}}{100} \qquad \text{czyli} \quad 0{,}14 \pm 0{,}12$$

5.6. Test istotności dla wariancji

Wariancja jako miara rozrzutu bywa wykorzystywana do oceny stopnia jednorodności bądź powtarzalności wyników eksperymentów. Test wariancji wymaga normalnego rozkładu populacji, z której wylosowano n-elementową próbę. Hipotezie zerowej mówiącej, że wariancja w populacji jest równa pewnej wartości σ_0^2 $\left(H_0 : \sigma^2 = \sigma_0^2\right)$ przeciwstawiamy na ogół hipotezę alternatywną $H_1 : \sigma^2 > \sigma_0^2$. Stosujemy więc test jednostronny, co jest spowodowane tym, że zwykle jedynie wariancja większa od pewnego progu jest niekorzystna. Do weryfikacji hipotezy budujemy statystykę χ^2

$$\chi^2 = \frac{(n-1)\,s^2}{\sigma_0^2} \tag{5.21}$$

gdzie: s^2 — oszacowanie wariancji z próby.
Statystyka (5.21) ma rozkład χ^2 o $n-1$ stopniach swobody. Tablice rozkładów χ^2 o k stopniach swobody podają wartości krytyczne ${}_{\alpha}\chi^2_{(k)}$ tych rozkładów właśnie dla jednostronnych obszarów krytycznych (patrz rys. 5.3))



Rys. 5.3 obszar krytyczny $P\left\{\chi^2 \geq {}_{\alpha}\chi^2\right\} = \alpha$ rozkładu χ^2.

$$P\left\{\chi_2 \geq {}_{\alpha}\chi^2_{(k)}\right\} = \alpha$$

Hipotezę H_0 odrzucamy przyjmując hipotezę alternatywną H_1 gdy $\chi^2 \geq {}_{\alpha}\chi^2_{(k)}$. W przeciwnym przypadku nie ma podstaw do odrzucenia hipotezy zerowej.

5.7 Porównywanie dwóch wariancji

Test porównania dwu wariancji wykonujemy najczęściej wówczas, gdy trzeba sprawdzić czy spełnione jest założenie równości wariancji w teście t (wzory (5.9) i (5.10)) dla dwóch średnich.

Zakładamy, że z dwóch populacji o rozkładach normalnych pobrano próby o liczebnościach odpowiednio n_1 i n_2 elementów. Sprawdzamy hipotezę zerową, że wariancje w obu populacjach oznaczone σ_1^2 i σ_2^2 są równe $\left(H_0 : \sigma_1^2 = \sigma_2^2\right)$. Również tutaj wykonujemy test jednostronny i formułujemy hipotezę alternatywną jako $H_1 : \sigma_1^2 > \sigma_2^2$

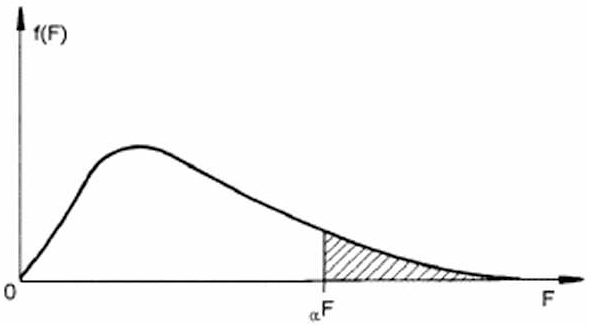
Równość dwóch wariancji powodowałaby, że ich stosunek σ_1^2/σ_2^2 byłby równy jedności. Właśnie stosunek uzyskanych z obu prób oszacowań wariancji s_1^2 i s_2^2 (przy czym $s_1^2 > s_2^2$)

$$F = \frac{s_1^2}{s_2^2} \tag{5.22}$$

jest statystyką wykorzystywaną podczas weryfikacji hipotezy o równości wariancji. Zmienna losowa F ze wzoru (5.22) charakteryzuje się jednym z częściej wykorzystywanych rozkładów,

tzw. rozkładem F Snedecora o $n_1 - 1$ stopniach swobody licznika i $n_2 - 1$ stopniach swobody mianownika. Tablice rozkładu F przy k i m stopniach swobody podają wartości krytyczne $_{\alpha}F_{(m)}^{(k)}$ tych rozkładów dla prawostronnych obszarów krytycznych (patrz rys. 5.4)

$$P\left\{F > {}_{\alpha}F_{(m)}^{(k)}\right\} = \alpha$$



Rys. 5.4 Obszar krytyczny $P\{F \geq {}_{\alpha}F\} = \alpha$ rozkładu F.

Hipotezę H_0 odrzucamy przyjmując hipotezę alternatywną H_1 gdy obliczone według (5.22) F spełnia zależność

$$F \geq {}_{\alpha}F_{(n_2-1)}^{(n_1-1)}$$

W przeciwnym przypadku nie znajdujemy podstaw do odrzucenia hipotezy zerowej.

Przykład 5.5 (według [Greń])

Zmierzono w dwóch ulach średnicę komórek plastra zbudowanego przez pszczoły. Dla 7 wylosowanych komórek z pierwszego plastra otrzymano następujące wyniki (w ulu): 5,36; 5,20; 5,28; 5,16; 5,30; 5,08; 5,23; analogicznie dla drugiego ula otrzymano: 5,15; 5,04; 5,30; 5,22; 5,19; 5,24; 5,12. Na poziomie istotności $\alpha = 0{,}05$ zweryfikować hipotezę, że średnie długości średnic komórek w plastrach pochodzących z dwu różnych uli są równe. Najpierw sprawdzimy, czy można uważać, że wariancje średnic komórek plastra w obu płatach są równe. Mamy:

$n_1 = n_2 = 7$

$\overline{x}_1 = 5{,}23$

$\overline{x}_2 = 5{,}18$

$s_1^2 = 0{,}008767$

$s_2^2 = 0{,}0073$

$F = 0{,}008767 / 0{,}0073 = 1{,}201$

$_{0{,}05}F_{(6)}^{(6)} = 4{,}28$

Ponieważ $F < {}_{\alpha}F_{(n_2-1)}^{(n_1-1)}$, więc nie ma żadnych podstaw do odrzucenia hipotezy o równości wariancji w obu populacjach średnic komórek plastra. Stosujemy więc test t. Ze wzoru (5.10) mamy

$$s^2 = 0{,}00803$$

i dalej ze wzoru (5.9)

$$t = \frac{5{,}23 - 5{,}18}{\sqrt{0{,}002295}}$$

Ponieważ

$${}_{0{,}05}t_{(12)} = 2{,}179$$

nie znajdujemy podstaw do odrzucenia stwierdzenia o równych średnich średnicach komórek plastrów w obu ulach.

6. ANALIZA DANYCH JAKOŚCIOWYCH. WIELOPOLOWE TABLICE KONTYNGENCJI

6.1 Tablice czteropolowe

W badaniach biologicznych i medycznych bardzo często do tej samej populacji stosujemy dwie różne klasyfikacje o charakterze **jakościowym**. W niniejszym rozdziale omówimy sposoby wnioskowania statystycznego w takich przypadkach — w szczególności będziemy starali się odpowiedzieć na pytanie, czy istnieje jakiś związek, jakaś zależność między jedną klasyfikacją a drugą. Najpierw będzie rozważany najprostszy przypadek, kiedy mamy do czynienia z dwiema klasyfikacjami, z których każda dotyczy dwóch wzajemnie wykluczających się kategorii.

Każdy badany obiekt będzie więc należał do jednej z dwóch możliwych kategorii w ramach klasyfikacji pierwszej i równocześnie do jednej z dwóch możliwych kategorii w ramach drugiej klasyfikacji. Liczebności tak pogrupowanych obiektów można zapisać w tabeli o dwóch wierszach i dwóch kolumnach (zwanej tabelą czteropolową lub tabelą kontyngencji 2×2), tak jak to pokazuje poniższy przykład.

Przykład 6.1 [Parker]

Jako część studiów nad jednostronnością u człowieka badano związek między prawo- i leworęcznością a dominacją prawo- i lewooczną w grupie 400 dzieci w wieku szkolnym. Każde dziecko zakwalifikowano do prawo- lub leworęcznych i równocześnie do prawo- lub lewoocznych. Uzyskane wyniki przedstawiono w tabeli 6.1. W boczku i główce tabeli uwidoczniono tzw. sumy marginalne, czyli sumaryczne liczebności obiektów należących do odpowiednich kategorii w ramach każdej z klasyfikacji.

Tabela 6.1

Wyniki badań nad jednostronnością u młodzieży szkolnej

		Dominacja ręki		(Razem)
		Leworęczni	Praworęczni	
Dominacja oka	Lewooczni	27	110	137
	Prawooczni	27	236	263
(Razem)		54	346	400

6.1.1. Test niezależności χ^2

Zadaniem naszym będzie weryfikacja hipotezy zerowej, mówiącej że dwie badane klasyfikacje są wzajemnie niezależne. Warunkiem tej niezależności jest, aby dla każdego z czterech pól tabeli prawdopodobieństwo zakwalifikowania obiektu do tego pola było równe iloczynowi prawdopodobieństw zakwalifikowania obiektu do odpowiednich (wyznaczonych przez to pole) kategorii w każdej klasyfikacji z osobna. Warunek powyższy pozwala obliczyć takie liczebności oczekiwane we wszystkich polach tabeli, które wystąpiłyby, gdyby rzeczywiście niezależność obu klasyfikacji miała miejsce. Do obliczenia oczekiwanych liczebności są niezbędne sumy marginalne. Liczebności oczekiwane (oznaczane literą E) oblicza się jako:

$$E = \frac{\text{suma\_wiersza} \cdot \text{suma\_kolumny}}{\text{liczebność\_całkowita}} \tag{6.1}$$

Oznaczamy liczebności faktycznie zaobserwowane jako O. Im większe są różnice $O - E$ w każdym polu tabeli, tym więcej przemawia przeciwko hipotezie zerowej o niezależności obu klasyfikacji. Uzasadnione jest więc zbudowanie testu w oparciu o takie różnice. Nie można jednakże różnic tych bezpośrednio zsumować po wszystkich polach tabeli, gdyż zawsze różnice te zniosłyby się wzajemnie i otrzymalibyśmy w wyniku wartość zerową. Dlatego też oblicza się statystykę χ^2 zdefiniowaną jako:

$$\chi^2 = \sum \frac{(O-E)^2}{E} \tag{6.2}$$

przy czym sumowanie przebiega po wszystkich czterech polach tabeli. Składnikami χ^2 są kwadraty odchyleń wartości obserwowanych od oczekiwanych odniesione do wartości

oczekiwanych. Jest to zrozumiałe, gdyż np. odchylenie równe 5 ma większą wagę (znaczenie), gdy wartość oczekiwana wynosi 10 niż gdy $E = 1000$. Poszczególne składniki χ^2 są więc względnymi miarami niezgodności liczebności obserwowanych z hipotezą zerową dla poszczególnych pól tabeli. Obliczona zgodnie z (6.2) statystyka χ^2 ma przy założeniu prawdziwości hipotezy zerowej asymptotyczny rozkład χ^2 o jednym stopniu swobody. Hipotezę zerową o braku związku między dwoma klasyfikacjami odrzucamy, gdy

$$\chi^2 \geq {}_{\alpha}\chi^2_{(1)}$$

gdzie: ${}_{\alpha}\chi^2_{(1)}$ jest wartością krytyczną rozkładu χ^2 o jednym stopniu swobody dla poziomu istotności α. W przeciwnym przypadku nie ma podstaw do odrzucenia hipotezy zerowej.

Przykład 6.1 (kontynuacja)

Tabela 6.2

Analiza wyników badań nad jednostronnością u młodzieży szkolnej

			Dominacja ręki		(Razem)
			Leworęczni	Praworęczni	
Dominacja oka	Lewooczni	*O*	27	110	137
		E	18,5	118,5	
		O – *E*	8,5	–8,5	
	Prawooczni	*O*	27	236	263
		E	35,5	227,5	
		O – *E*	–8,5	8,5	
(Razem)			54	346	400

Tabela 6.2 przedstawia obliczone wartości oczekiwane oraz odchylenia wartości obserwowanych od oczekiwanych dla wyników badań nad jednostronnością u młodzieży szkolnej. Sposób obliczenia liczebności oczekiwanej dla np. lewego górnego pola w tabeli:

$$E = \frac{54}{400} \cdot 137 = 18{,}5$$

można wytłumaczyć w następujący sposób: jeżeli pomnożymy całkowity udział leworęcznych w próbie (równy 54/400) przez całkowitą liczebność lewoocznych (równą 137) to uzyskamy spodziewaną liczebność leworęcznych i równocześnie lewoocznych, przy

zachowaniu niezależności obu klasyfikacji. Tę niezależność można interpretować np. jako taki sam udział leworęcznych wśród zarówno lewoocznych jak i prawoocznych.

Obliczona według (6.2) wartość χ^2 wynosi:

$$\chi^2 = \frac{8{,}5^2}{18{,}5} + \frac{(-8{,}5)^2}{118{,}5} + \frac{(-8{,}5)^2}{35{,}5} + \frac{(8{,}5)^2}{227{,}5} = 6{,}866$$

Z tablic odczytujemy wartość krytyczną rozkładu χ^2 o jednym stopniu swobody przy $\alpha = 0{,}01$

$$_{0{,}01}\chi^2_{(1)} = 6{,}63$$

Ponieważ χ^2 jest większe od wartości krytycznej, więc wnioskujemy, że związek między dominacją ręki i dominacją oka jest istotny.

Podamy teraz praktyczne wzory do obliczenia χ^2 dla tabel czteropolowych. Jeżeli tabela zawiera wartości oznaczone w następujący sposób:

	1	2	
1	a	b	r_1
2	c	d	r_2
	s_1	s_2	N

to χ^2 wyznaczamy według wzoru:

$$\chi^2 = \frac{(ad - bc)^2}{r_1 r_2 s_1 s_2} \tag{6.3}$$

Dla tablic czteropolowych, zwłaszcza gdy liczebności oczekiwane są małe, stosujemy skorygowany wzór na χ^2 zawierający tzw. poprawkę na nieciągłość (Yatesa). Poprawka ta polega na odjęciu wartości 0,5 od bezwzględnej wartości różnicy $(O - E)$. Odpowiednie wzory mają postać:

$$\chi^2_c = \sum \frac{(|O - E| - \frac{1}{2})^2}{E} \tag{6.4}$$

lub

$$\chi^2_c = \sum \frac{(|ad - bc| - \frac{1}{2}N)^2 N}{r_1 r_2 s_1 s_2} \tag{6.5}$$

Jeżeli którakolwiek z liczebności oczekiwanych jest mniejsza niż 5 i $20 < N < 40$ lub jeśli $N < 20$ — to wtedy nie powinniśmy stosować testu niezależności opartego na χ^2, ale dokładny test Fishera, opisany w następnym punkcie.

Przykład 6.1 (dokończenie)

Zastosowanie poprawki Yatesa dla przykładu badań nad jednostronnością młodzieży daje skorygowaną wartość $\chi_c^2 = 6{,}09$. Wartość χ_c^2 jest nieco mniejsza od wartości obliczonej poprzednio bez stosowania korekcji, jednak związek między dwoma klasyfikacjami jest nadal istotny, tym razem na poziomie $\alpha = 0{,}025$, gdyż $_{0{,}025}\chi^2_{(1)} = 5{,}02$.

6.1.2. Dokładny test Fishera.

Jeżeli sumaryczna liczebność obserwacji jest mała lub jeśli zbyt małe są liczebności oczekiwane, wtedy stosujemy dokładny test dla tablic czteropolowych zwany w literaturze testem Fishera. Opiera on się na tym, że przy założeniu prawdziwości hipotezy zerowej dokładne prawdopodobieństwo otrzymania tabeli o liczebnościach obserwowanych

a	b	r_1
c	d	r_2
s_1	s_2	N

(6.6)

dane jest wzorem:

$$P = \frac{r_1!\, r_2!\, s_1!\, s_2!}{N!\, a!\, b!\, c!\, d!} \tag{6.7}$$

Można więc obliczyć prawdopobieństwa otrzymania wszystkich tabel czteropolowych o liczebnościach brzegowych takich, jakie występują w tabeli obserwowanej, przez co uzyska się dyskretny rozkład prawdopodobieństwa wszystkich możliwych tabel. Sumując prawdopodobieństwo wystąpienia tabeli obserwowanej i innych jeszcze mniej prawdopodobnych tabel usytuowanych w tym samym skrzydle rozkładu można ocenić poziom istotności α przy jakim ewentualnie dałoby się odrzucić hipotezę zerową. Sposób postępowania zostanie dokładniej przedstawiony na przykładzie liczbowym.

Przykład 6.2 (według [Armitage])

Badając wpływ sposobu karmienia niemowląt na stan ich uzębienia uzyskano wyniki przedstawione w tabeli 6.3. Ponieważ aż dwie liczebności oczekiwane są małe (mniejsze od 5), więc zachodzi potrzeba zastosowania dokładnego testu Fishera. Tworzymy wszystkie możliwe tabele 2×2 o liczebnościach brzegowych takich, jak w tabeli obserwowanej.

Tabela 6.3

Dane dotyczące stanu uzębienia niemowląt w związku ze sposobem ich karmienia

			Niemowlęta z uzębieniem		(Razem)
			normalnym	wadliwym	
Karmione	piersią	*O* *E*	4 (2,4)	16 (17,6)	20
	z butelki	*O* *E*	1 (2,6)	21 (19,4)	22
(Razem)			5	37	42

Takich tabel jest sześć poczynając od tabeli mającej wartość „*a*" w lewym górnym rogu równą 0, a kończąc na tabeli mającej $a = 5$. Wszystkie możliwe tablice 2×2 pokazano w tabeli 6.4. Prawdopodobieństwo P_0 wystąpienia tabeli, w której $a = 0$ można obliczyć ze wzoru:

Tabela 6.4

Wszystkie możliwe tabele kontyngencji 2×2 o liczebnościach brzegowych identycznych jak w tabeli 6.3

0	20	20
5	17	22
5	37	42

$a = 0$

1	19	20
4	18	22
5	37	42

$a = 1$

2	18	20
3	19	22
5	37	42

$a = 2$

3	17	20
2	20	22
5	37	42

$a = 3$

4	16	20
1	21	22
5	37	42

$a = 4$

5	15	20
0	22	22
5	37	42

$a = 5$

$$P_0 = \frac{(c_0 + d_0)!\,(b_0 + d_0)!}{N!\,d_0!} \tag{6.8}$$

gdyż do takiej postaci redukuje się w tym przypadku wzór (6.7). Prawdopodobieństwa dla pozostałych tabel wyznacza się ze wzoru rekurencyjnego:

$$P_{k+1} = P_k \frac{b_k c_k}{(a_k + 1)(d_k + 1)} \tag{6.9}$$

przy czym a_k, b_k, c_k, d_k oznaczają wartości w odpowiednich polach tabeli dla $a = k$. Mamy więc

$$P_0 = \frac{22!\,37!}{42!\,17!} = 0{,}03096$$

$$P_1 = \frac{5 \cdot 20}{1 \cdot 18} \cdot P_0 = 0{,}1720, \qquad \text{itd.}$$

Ostatecznie otrzymujemy następujący rozkład prawdopodobieństwa tabel kontyngencji:

$P_0 = 0{,}0310$
$P_1 = 0{,}1720$
$P_2 = 0{,}3440$
$P_3 = 0{,}3096$
$\boxed{P_4 = 0{,}1253}$
$P_5 = 0{,}0182$
suma 1,0001

Wielkość P_4 jest prawdopodobieństwem wystąpienia tabeli obserwowanej. Dla potrzeb testu jednostronnego (który można stosować, gdy z góry jesteśmy w stanie przewidzieć kierunek odchylenia od tabeli oczekiwanej) sumujemy prawdopodobieństwo tabeli obserwowanej i pozostałych jeszcze mniej prawdopodobnych tabel o tym samym kierunku odchylenia od centrum rozkładu

$$P = P_4 + P_5 = 0{,}1253 + 0{,}0182 = 0{,}1435$$

Stosując test jednostronny moglibyśmy więc odrzucić hipotezę zerową dopiero przy poziomie istotności $\alpha = P = 0{,}1435$. Oczywiście taki poziom istotności jest nie do przyjęcia, stąd stwierdzamy, że nie ma podstaw do odrzucenia hipotezy o braku zależności między stanem uzębienia niemowląt a sposobem ich karmienia.

Na ogół nie można jednak stosować testu jednostronnego i wówczas poziom istotności równa się podwójnej sumie P obliczonej poprzednio

$$\alpha = 2P$$

W naszym przypadku poziom istotności wyniósłby

$$\alpha = 2 \cdot 0{,}1435 = 0{,}2870$$

co stanowi wielkość już zupełnie nie kwalifikującą się do przyjęcia. Niektórzy autorzy zamiast podwajać liczbę P proponują dodać do niej prawdopodobieństwa wystąpienia tabel, które są co najmniej tak odległe od oczekiwanej (w sensie równie małego prawdopodobieństwa wystąpienia), ale są położone po drugiej stronie centrum rozkładu. U nas dałoby to poziom istotności

$$\alpha = P' = P + P_0 = 0{,}1435 + 0{,}0310 = 0{,}1745$$

6.1.3. Miary siły związku

Przykład 6.3 (na podstawie [Parker])

Oceniano stopień zarastania trawnika przez stokrotki (Bellis perennis). Ocenę wykonano na 40 losowo wybranych kwadratach, na których zarejestrowano „obecność" lub „nieobecność" stokrotek. Zanotowano łącznie 16 „obecności". Następnie na omawianym trawniku dokonano oprysku selektywnie działającymi herbicydami i po trzech tygodniach wykonano powtórną ocenę pokrycia terenu przez stokrotki. Tym razem, również na 40 losowo wybranych kwadratach zanotowano łącznie 12 „obecności". Wykorzystując test χ^2 Pearsona zbadać istotność spadku stopnia pokrycia terenu przez stokrotki w wyniku zastosowanego oprysku.

Tablica kontyngencji 2 × 2 dla tego przypadku została przedstawiona w tabeli 6.5. Obliczona w/g wzoru (6.5) wartość χ^2_c wynosi

Tabela 6.5

Wyniki badań stopnia pokrywania terenu przez stokrotki w powiązaniu z działaniem herbicydu

	Przed opryskiem	Po oprysku	(Razem)
Stokrotki „obecne"	16	12	28
Stokrotki „nieobecne"	24	28	52
(Razem)	40	40	80

$$\chi_c^2 = 0{,}8357$$

co oczywiście nie jest wynikiem uprawniającym do odrzucenia hipotezy zerowej o braku związku pomiędzy stopniem pokrycia terenu przez stokrotki a działaniem herbicydu.

Jednakże gdybyśmy dysponowali np. dziesięciokrotnie większą próbą, to wówczas tablica kontyngencji zachowująca te same proporcję rozkładu, co poprzednio miałaby postać:

160	120	280
240	280	520
400	400	800

a obliczona na jej podstawie wartość χ_c^2 równa

$$\chi_c^2 = 8{,}357$$

byłaby dziesięciokrotnie większa niż poprzednia i wskazywałaby na wysoką istotność badanego związku (na poziomie istotności $\alpha = 0{,}01$).

Jak widać z tego przykładu, istnienie związku i siła związku to zagadnienia w dużej mierze od siebie niezależne. Test χ^2 i poziom istotności informują tylko o prawdopodobieństwie istnienia związku, a nie o jego natężeniu. Istotność związku statystycznego zależy także od wielkości badanej próby. Gdy próba jest duża, bardzo łatwo można wykazać istotność statystyczną nawet bardzo słabego związku. W przypadku małej próby związek musi być znacznie silniejszy, aby mógł być uznany za istotny. Wobec tego w wielu wypadkach ważna jest odpowiedź na pytanie: jeśli związek istnieje, to jakie jest jego natężenie?

Testując obliczoną statystykę χ^2 uzyskujemy odpowiedź na pytanie: czy można uznać, że związek istnieje? Do pomiaru siły związku wartość χ^2 bezpośrednio się nie nadaje, gdyż, jak widzieliśmy w przykładzie 6.3, zależy ona od N i rośnie wraz ze wzrostem próby. Najbardziej naturalnym miernikiem siły związku jest wielkość Φ określona jako

$$\Phi^2 = \frac{\chi^2}{N} \tag{6.10}$$

Praktycznie Φ obliczamy dla tablic 2×2 ze wzoru:

$$\Phi = \sqrt{\frac{(ad - bc)^2}{r_1 r_2 s_1 s_2}} \tag{6.11}$$

Oprócz Φ stosujemy również standaryzowaną miarę Pearsona r_p określaną jako:

$$r_p = \sqrt{\frac{2\,(ad - bc)^2}{r_1\, r_2\, s_1\, s_2 + (ad - bc)^2}} \tag{6.12}$$

lub miarę Q Kendalla:

$$Q = \frac{ad - bc}{ad + bc} \tag{6.13}$$

Mierniki określone wzorami (6.11) i (6.12) przyjmują wartości z przedziału <0, 1>, zaś miernik Q z przedziału <–1, +1>. Wartość zerowa wskazuje na zupełny brak związku, zaś 1 (czy ±1) na związek o maksymalnej sile. Tabela 6.6 podaje wartości mierników dla kilku przykładowych tablic czteropolowych. Należy zauważyć, że znak miernika Q Kendalla oznacza „dominację" określonej przekątnej. Miernik ten, w odróżnieniu od pozostałych, ma jeszcze jedną istotną własność, a mianowicie traktuje on tablicę czteropolową w której wystąpiło choć jedno zero, za reprezentantkę „pełnego" związku statystycznego między klasyfikacją kolumn i klasyfikacją wierszy. Decyzja o wyborze któregoś z mierników siły związku zależy w dużej mierze od celu badania statystycznego. Więcej na ten temat można przeczytać u Blalocka [Blalock].

Tabela 6.6

Wartości miar siły związku dla wybranych tablic kontyngencji 2 × 2

Tablica			Φ	Q	r_p
50	0	50	1	1	1
0	50	50			
50	50	100			
40	0	40	0,82	1	0,89
10	50	60			
50	50	100			
40	10	50	0,60	0,88	0,73
40	10	50			
50	50	100			

Tablica			Φ	Q	r_p
20	50	70	0,65	–1	0,77
30	0	30			
50	50	100			
20	30	50	0,20	–0,38	0,28
30	20	50			
50	50	100			
25	25	50	0	0	0
25	25	50			
50	50	100			

6.1.4 Test interakcji

Poprzednio (patrz punkty 6.1.1 i 6.1.2) weryfikując hipotezę zerową o braku zależności między dwoma klasyfikacjami w tablicy czteropolowej — ustalaliśmy wartości oczekiwane i obliczaliśmy statystykę χ^2 w oparciu o niezmieniane sumy marginalne. Dysponowaliśmy wówczas jednym stopniem swobody (jeżeli znamy sumy marginalne to wystarczy znać wartość liczebności w jednym polu tabeli, aby odtworzyć całą tabelę kontyngencji). Jeżeli chcemy bardziej dokładnie zbadać zależności wzajemne w tablicy czteropolowej i nie korzystać z sum marginalnych, to mamy do dyspozycji 3 stopnie swobody (znając sumaryczną liczebność próby i trzy liczebności w poszczególnych polach, czwartą brakującą możemy obliczyć przez odejmowanie). Podamy poniżej przykład testowania hipotez nie wymagających korzystania z sum marginalnych.

Przykład 6.4 (według [Parker])

Podwójnie heterozygotyczne rośliny kukurydzy (Zea mays) zostały skrzyżowane wstecznie z roślinami podwójnie recesywnymi. Otrzymano cztery rodzaje potomstwa:

czerwona owocnia i niekarłowate	PD	— 204 rośliny
czerwona owocnia i karłowate	Pd	— 153 rośliny
bezbarwna owocnia i niekarłowate	pD	— 154 rośliny
bezbarwna owocnia i karłowate	pd	— 165 roślin
Łącznie		676 roślin

Spodziewamy się, że obie klasyfikacje $P - p$ i $D - d$ będą niezależne i liczebności w każdym polu tablicy czteropolowej będą równe. Stąd wszystkie liczebności oczekiwane będą wynosić:

$$E = \frac{676}{4} = 169$$

Pełną tablicę kontyngencji przedstawia tabela 6.7.

Tabela 6.7

Wyniki badań krzyżówek kukurydzy *(Zea mays)*

		P	*p*	(Razem)
D	O	204	154	358
	E	169	169	338
	O – E	35	–15	20
d	O	153	165	318
	E	169	169	338
	O – E	–16	–4	–20
(Razem)		357	319	676
		338	338	
		19	–10	

W przypadku, gdy wszystkie oczekiwane liczebności są równe, wzór (6.2) można zapisać jako

$$\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{\sum O^2}{E} - N \qquad (6.14)$$

Wykorzystując powyższe obliczamy sumaryczną wartość χ^2

$$\chi^2 = \frac{204^2 + 153^2 + 154^2 + 165^2}{169} - 676 = 10{,}189$$

która porównana z wartością krytyczną przy trzech stopniach swobody i poziomie istotności 0,05

$$_{0,05}\chi^2_{(3)} = 7{,}81$$

wskazuje, że odchylenia od wartości oczekiwanych są istotne. Nie wiemy czy jest to spowodowane odchyleniem od oczekiwanego stosunku 1:1 w klasyfikacji $P - p$, czy odchyleniem od takiegoż samego spodziewanego stosunku 1:1 w klasyfikacji $D - d$, czy

też tym, że klasyfikacje nie są niezależne, to znaczy istnieje sprzężenie (interakcja) między P i D oraz p i d. Obliczamy składniki sumarycznego χ^2 odpowiadające tym trzem przypadkom (oznaczenia jak w (6.6)).

$$\chi^2_{P-p} = \frac{[(a+c)-(b+d)]^2}{N} = 2{,}136 \tag{6.15}$$

$$\chi^2_{D-d} = \frac{[(a+b)-(c+d)]^2}{N} = 2{,}366 \tag{6.16}$$

$$\chi^2_{inter} = \frac{[(a+d)-(b+c)]^2}{N} = 5{,}686 \tag{6.17}$$

wszystkie powyższe statystyki χ^2 charakteryzują się jednym stopniem swobody. Zachodzi tu

$$\chi^2 = \chi^2_{P-p} + \chi^2_{D-d} + \chi^2_{inter}$$

Wiedząc, że wartość krytyczna χ^2 przy jednym stopniu swobody wynosi

$$_{0{,}05}\chi^2_{(1)} = 3{,}84$$

i porównując ją z wartościami obliczonych składników całkowitego χ^2 widzimy, że znamienność wykryta dla sumarycznego χ^2 ma swoje źródło w istotności sprzężenia (interakcji) między P i D oraz p i d, natomiast odchylenia od oczekiwanego stosunku 1:1 nie są istotne dla żadnej z klasyfikacji. Wykryta istotna interakcja ma swoje źródło w znacznie przekraczającej wartość oczekiwaną liczebności potomstwa typu PD w lewym górnym polu tablicy (por. tabela 6.7).

Dla testowania „mendlowskiego" stosunku 9:3:3:1 dla składników χ^2 (por. zależności (6.15)-(6.17)) stosuje się wzory (oznaczenia jak poprzednio)

$$\chi^2_{P-p} = \frac{1}{3N}[(a+c)-3(b+d)]^2 \tag{6.18}$$

$$\chi^2_{D-d} = \frac{1}{3N}[(a+b)-3(c+d)]^2 \tag{6.19}$$

$$\chi^2_{inter} = \frac{1}{9N}\left[(a + 9d) - 3\,(b + c)\right]^2 \tag{6.20}$$

6.2 Tablice kontyngencji 2x*k*

6.2.1 Porównanie kilku częstości

Przykład 6.5 (według [Armitage])

Powróćmy do danych z przykładu 5.2, gdzie porównywano dwie częstości względne. Tamte dane (przypomnijmy: spośród 257 pacjentów leczonych metodą *A* zmarło 41, a spośród 244 pacjentów leczonych metodą *B* zmarło 64) można przedstawić przy pomocy tablicy kontyngencji 2 × 2, jak to pokazuje tabela 6.8. Obliczona dla tej tablicy wartość χ^2 wynosi

$$\chi^2 = 7{,}978$$

Tabela 6.8

Tablica przedstawiająca wyniki próby klinicznej dwóch metod leczenia A i B

		Leczenie		(Razem)
		A	B	
Wynik	zgon	41	64	105
	przeżycie	216	180	396
(Razem)		257	244	501

Wiadomo, że rozkład $\chi^2_{(1)}$ jest rozkładem kwadratu zmiennej losowej o standaryzowanym rozkładzie normalnym. W przykładzie 5.2 dla porównania dwóch frakcji obliczono wartość statystyki u mającej standaryzowany rozkład normalny. Wartość ta była równa –2.82, co podniesione do kwadratu daje $(-2.82)^2 = 7{,}95$ czyli wartość równą (z dokładnością do błędu zaokrągleń) wartości χ^2.

Pokazaliśmy, że test niezależności χ^2 dla tablic czteropolowych może być stosowany dla porównywania dwóch częstości. Podobnie stosując test χ^2 dla tablic kontyngencji 2 × *k* możemy porównywać k częstości. Załóżmy, że dysponujemy wynikami badań *k*

różnych grup, w każdej i-tej grupie w próbie o liczebności n_i znajdujemy r_i obserwacji charakteryzujących się cechą A. Przyjmijmy oznaczenia:

Grupa	1	2	$\ldots i \ldots$	k
cecha A	r_1	r_2	$\ldots r_i \ldots$	r_k
brak cechy A	$n_1 - r_1$	$n_2 - r_2$	$\ldots n_i - r_i \ldots$	$n_k - r_k$
Ogółem	n_1	n_2	n_i	n_k
Frakcja obserwacji z cechą A	p_1	p_2	$\ldots p_i = \frac{r_i}{n_i} \ldots$	p_k

Stawiamy hipotezę zerową, że wszystkie k próbek uzyskano losowo z populacji o tej samej frakcji wyników odznaczających się cechą A. Obliczamy wartość statystyki χ^2 według znanego wzoru

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

(sumowanie przebiega po wszystkich $2k$ polach tablicy) przy czym wartości oczekiwane wyznacza się z zależności

$$E = \frac{\text{suma\_wiersza} \cdot \text{suma\_kolumny}}{\text{liczebność\_całkowita}}$$

Można również korzystać ze wzoru równoważnego lecz wygodniejszego do obliczeń

$$\chi^2 = \frac{\sum_{i=1}^{k} \frac{r_i^2}{n_i} - \frac{R^2}{N}}{P(1-P)} \tag{6.21}$$

Tak obliczoną wartość χ^2 porównujemy z wartością krytyczną dla ustalonego poziomu istotności α odczytaną z tablic rozkładu χ^2 o $k - 1$ stopniach swobody. Hipotezę zerową odrzucamy gdy

$$\chi^2 \geq \alpha \chi^2_{(k-1)}$$

zaś w przeciwnym przypadku nie ma podstaw do odrzucenia hipotezy zerowej.

Przykład 6.6 (według [Armitage])

Wyniki badań dzieci ze względu na nosicielstwo bakterii *Streptococcus pyogenes* przy uwzględnieniu wielkości migdałków przedstawia tabela 6.9. Celem jest porównanie frakcji nosicieli bakterii w każdej z trzech grup dzieci: z migdałkami normalnej wielkości, z migdałkami powiększonymi oraz z migdałkami bardzo powiększonymi. Stawiamy hipotezę zerową, że „prawdziwe" frakcje nosicieli są we wszystkich trzech grupach równe.

Tabela 6.9

Liczba dzieci — nosicieli i nienosicieli bakterii** Streptococcus pyogenes **w zależności od wielkości migdałków

	Migdałki			(Razem)
	nie powiększone	powiększone	bardzo powiększone	
Nosiciele	19	29	24	72
Nienosiciele	497	560	269	1326
(Razem)	516	589	293	1398
Frakcja nosicieli	0,0368	0,0492	0,0892	(0,0515)

Obliczenie według wzoru (6.21) wartości statystyki χ^2 daje wartość

$$\chi^2 = 7{,}88$$

co w porównaniu z wartością krytyczną

$$_{0{,}05}\chi^2_{(2)} = 5{,}99$$

pozwala odrzucić hipotezę zerową i stwierdzić, że przy poziomie istotności $\alpha = 0{,}05$ różnice między frakcjami są znamienne. Dalszą analizę tych danych przedstawimy w następnym punkcie.

6.2.2 Test trendu częstości

Jeżeli w tablicy kontyngencji $2 \times k$ k grup ułożonych jest w pewnym naturalnym porządku, to w przypadku uzasadnionym można pokusić się o przetestowanie istnienia znamiennego trendu frakcji (częstości) od grupy 1 do grupy k. Test trendu wymaga, aby poszczególne grupy mogły być uporządkowane według pewnego jasno określonego kryterium. Wtedy możemy poszczególnym grupom przypisać wartości pewnej zmiennej ilościowej x. I tak jeżeli grupami są np. przedziały wiekowe pacjentów, to wartościami zmiennej x mogą być liczby x_i będące środkami przedziałów. Jeżeli grupy mają charakter typowo jakościowy — wartościami x_i mogą być kolejne liczby całkowite. Dla potrzeb testu trendu oblicza się statystykę χ_1^2 zgodnie ze wzorem:

$$\chi_1^2 = \frac{N\left(N\sum_{i=1}^{k} r_i x_i - R\sum_{i=1}^{k} n_i x_i\right)}{R\,(N-R)\left[N\sum_{i=1}^{k} n_i x_i^2 - \left(\sum_{i=1}^{k} n_i x_i^2\right)^2\right]} \tag{6.22}$$

gdzie x_i jest wartością zmiennej x w i-tej grupie. Statystyka ta, mająca w przybliżeniu rozkład χ^2 o jednym stopniu swobody jest częścią całkowitej wartości χ^2 (liczonej w/g wzoru ogólnego (6.2) lub (6.21)) „odpowiedzialną" za występowanie liniowego trendu frakcji p_i względem zmiennej x_i. Wartość χ_1^2 większa od wartości krytycznej dla danego poziomu istotności pozwala na wnioskowanie o istnieniu znamiennego trendu frakcji. Różnicę między całkowitą wartością χ^2 a obliczoną wartością χ_1^2 oznaczamy przez χ_2^2

$$\chi_2^2 = \chi^2 - \chi_1^2 \tag{6.23}$$

χ_2^2 ma w przybliżeniu rozkład χ^2 o $k-2$ stopniach swobody i jest miarą odchyleń poszczególnych frakcji od ogólnego trendu liniowego. Wnioskowanie na podstawie statystyki χ_2^2 prowadzimy według ogólnych zasad.

Przykład 6.6 (kontynuacja z punktu 6.2.1)

Kontynując analizę danych zawartych w tabeli 6.9 można próbować odpowiedzieć na pytanie, czy udział nosicieli bakterii wykazuje tendencję wzrostową wraz z powiększaniem się migdałków. Ze względu na brak możliwości precyzyjnego przyporządkowania wartości zmiennej ilościowej x wielkościom migdałków, przyjmujemy zapis:

migdałki niepowiększone $x_1 = -1$
migdałki powiększone $x_2 = 0$
migdałki bardzo powiększone $x_3 = 1$

Wykorzystując wzory (6.22) i (6.23) mamy

$$\chi_1^2 = \frac{1398\,[1398 \cdot 5 - 72\,(-233)]^2}{72 \cdot 1326\,[1398 \cdot 809 - (-233)^2]} = 7{,}19$$

$$\chi^2 = 7{,}88$$

$$\chi_2^2 = 7{,}88 - 7{,}19 = 0{,}69$$

Ponieważ

$${}_{0{,}05}\chi_{(1)}^2 = 3{,}84 \qquad {}_{0{,}01}\chi_{(1)}^2 = 6{,}63$$

więc

$$\chi_1^2 > {}_{0{,}01}\chi_{(1)}^2$$

$$\chi_2^2 < {}_{0{,}05}\chi_{(1)}^2 \qquad \text{(gdyż } k - 2 = 3 - 2 = 1\text{)}$$

Uzyskaliśmy silne potwierdzenie istotności trendu i nieistotności odchyleń od trendu.

6.2.3 Test χ^2 w klasyfikacji hierarchicznej

Przy pomocy tablic kontyngencji $2 \times k$ i wyodrębniania poszczególnych składników statystyki χ^2 można porównywać kilka różnych klasyfikacji tego samego materiału obserwacyjnego utworzonych przy zastosowaniu podejścia hierarchicznego. Sposób postępowania przedstawia poniższy przykład.

Przykład 6.7 (według [Armitage])

Wyniki doświadczenia poświęconego działaniu dwóch różnych środków owadobójczych na muchy przedstawia tabela 6.10. W doświadczeniu badano po dwie partie każdego środka i każdą z tych czterech partii stosowano dwukrotnie (dokonywano z każdą partią dwóch prób). Porównywano frakcje much padłych.

Ogólna frakcja much padłych wynosi

$$P = \frac{346}{391} = 0{,}8849$$

Tabela 6.10

Wyniki doświadczenia z dwoma środkami owadobójczymi

		Środek A				Środek B				(Razem)
		Partia A_1		Partia A_2		Partia B_1		Partia B_2		
		Próba 1	Próba 2	Próba 1	Próba 2	Próba 1	Próba 2	Próba 1	Próba 2	
Muchy	Padły	49	43	43	48	41	44	39	39	346
	Przeżyły	2	5	4	1	5	8	11	9	45
(Razem)		51	48	47	49	46	52	50	48	391

Obliczamy całkowitą wartość χ^2 korzystając z (6.21)

$$\chi^2 = \frac{\frac{2^2}{51} + \frac{5^2}{48} + \frac{4^2}{47} + \cdots + \frac{9^2}{48} - \frac{45^2}{391}}{0{,}8849 \cdot 0{,}1151} = \frac{1{,}6628}{0{,}1019} = 16{,}32$$

Porównując tę wartość z wartością krytyczną rozkładu χ^2 przy 7 stopniach swobody i poziomie istotności $\alpha = 0{,}05$

$$_{0{,}05}\chi^2_{(7)} = 14{,}07$$

mamy

$$\chi^2 > {}_{0{,}05}\chi^2_{(7)}$$

a więc wnioskujemy, że różnice między poszczególnymi frakcjami much padłych są istotne.

Chcąc sprawdzić, czy zaobserwowany wynik powodowany jest różnicami między poszczególnymi próbami, partiami czy środkami owadobójczymi, obliczamy składniki ogólnego χ^2 odpowiedzialne za te składowe zależności. Najpierw rozważymy zależności pomiędzy próbami w ramach partii wyznaczając wartość odpowiedniego składnika, który oznaczymy jako $\chi^2_{próby}$ i obliczymy z zależności:

$$\chi^2_{próba} = \frac{\sum\limits_{partie}\left[\sum\limits_{\substack{próby\\ w\ ramach\\ partii}} \frac{r_i^2}{n_i} - \frac{R^2}{N}\right]}{P(1-P)}$$

co odpowiada zdezagregowaniu tablicy pierwotnej, otrzymaniu układu 4 tablic składowych (jak to przedstawia tabela 6.11), a następnie zsumowaniu liczników odpowiednich wyrażeń χ^2 dla poszczególnych tablic, przy zachowaniu mianownika identycznego, jak dla tabeli pierwotnej.

Tabela 6.11

Dezagragacja tablicy kontyngencji przedstawionej w tabeli 6.10 dla potrzeb badania zależności pomiędzy próbami

Partia A_1

	1	2	
Padły	49	43	92
Przeżyły	2	5	7
	51	48	99

Partia A_2

	1	2	
Padły	43	48	91
Przeżyły	4	1	5
	47	49	96

Partia B_1

	1	2	
Padły	41	44	85
Przeżyły	5	8	13
	46	52	98

Partia B_1

	1	2	
Padły	39	39	78
Przeżyły	11	9	20
	50	48	98

Mamy:

$$\chi^2_{próby} = \frac{\left(\frac{2^2}{51} + \frac{5^2}{48} - \frac{7^2}{99}\right) + \left(\frac{4^2}{47} + \frac{1^2}{49} + \frac{5^2}{96}\right) + \cdots + \left(\frac{11^2}{50} + \frac{9^2}{48} + \frac{20^2}{98}\right)}{0{,}8849 \cdot 0{,}1151} = 2{,}75$$

$\chi^2_{próby}$ ma rozkład χ^2 o 4 stopniach swobody (liczba stopni swobody równa się tutaj sumie ilości stopni swobody tablic składowych). Ponieważ

$$_{0,05}\chi^2_{(4)} = 9{,}49$$

więc

$$\chi^2_{próby} < {}_{0,05}\chi^2_{(4)}$$

co świadczy o nieistotności różnic pomiędzy próbami.

Dla zbadania zależności pomiędzy partiami obliczamy składnik χ^2_{partie}

$$\chi^2_{próba} = \frac{\sum_{środki}\left[\sum_{\substack{próby\\ w\ ramach\\ partii}} \frac{r_i^2}{n_i} - \frac{R^2}{N}\right]}{P(1-P)}$$

co odpowiada zagregowaniu czterech tablic z tabeli 6.11, utworzeniu z ich sum brzegowych dwóch tablic pokazanych w tabeli 6.12 i utworzeniu na podobnych jak poprzednio zasadach „sumy" wyrażeń χ^2 dla obu tablic.

Tabela 6.12

Przekształcenie danych z tablicy kontyngencji przedstawionej w tabeli 6.10 dla potrzeb badania zależności pomiędzy partiami *(por. także tab. 6.11)*

Środek *A*

	A_1	A_2	
Padły	92	91	183
Przeżyły	7	5	12
	99	96	195

Środek *B*

	B_1	B_2	
Padły	85	78	163
Przeżyły	13	20	33
	98	98	196

Dokonując obliczeń otrzymujemy:

$$\chi^2_{partie} = \frac{\left(\frac{7^2}{99} + \frac{5^2}{96} - \frac{12^2}{195}\right) + \left(\frac{13^2}{98} + \frac{20^2}{98} + \frac{33^2}{196}\right)}{0{,}8849 \cdot 0{,}1151} = 2{,}62$$

χ^2_{partie} ma rozkład χ^2 o dwóch stopniach swobody. Ponieważ

$$_{0,05}\chi^2_{(2)} = 5{,}99$$

więc

$$\chi^2_{partie} < {}_{0,05}\chi^2_{(2)}$$

czyli również — rozważana w kontekście przeżycia much — zależność między partiami nie jest istotna.

Wreszcie dla zbadania zależności między środkami owadobójczymi obliczamy $\chi^2_{środki}$ według zależności:

$$\chi^2_{środki} = \frac{\sum_{środki} \frac{r_i^2}{n_i} - \frac{R^2}{N}}{P(1-P)}$$

co jest wynikiem dalszej hierarchicznej agregacji danych, przedstawionej w tabeli 6.13. $\chi^2_{środki}$ ma rozkład χ^2 o jednym stopniu swobody.

Tabela 6.13

Zagregowanie danych z tabeli 6.10 dla potrzeb porównania efektywności środków owadobójczych

	Środki		(Razem)
	A	B	
Muchy padły	183	163	346
Muchy przeżyły	12	33	45
(Razem)	195	196	391
Frakcja padłych	0,9385	0,8316	(0,8849)

Otrzymujemy:

$$\chi^2_{środki} = \frac{\frac{12^2}{183} + \frac{33^2}{196} - \frac{45^2}{391}}{0{,}8849 \cdot 0{,}1151} = 10{,}95$$

$$_{0,05}\chi^2_{(1)} = 3,84$$

$$\chi^2_{środki} > {}_{0,05}\chi^2_{(1)}$$

Uzyskaliśmy silne potwierdzenie istotności różnicy działania środków owadobójczych na korzyść środka *A*. Poza tym widzimy, że obliczone składniki χ^2 po zsumowaniu dają rzeczywiście wartość χ^2 dla testu ogólnego:

$$\chi^2_{próby} + \chi^2_{partie} + \chi^2_{środki} = 2,75 + 2,62 + 10,95 = 16,32 = \chi^2$$

6.2.4 Kombinowany test niejednorodności i zgodności

Poprzednio (por. punkt 6.2.1) omawiany test porównywania k częstości za pomocą tablicy kontyngencji $2 \times k$, czy też, gdyby powiedzieć to innymi słowami, test braku związku (niezależności) między dwoma klasyfikacjami (zakładającymi podział odpowiednio na dwie oraz k kategorii) wymagał takiego wyznaczania liczebności oczekiwanych, aby zachowane były sumy marginalne. Do tego potrzebowaliśmy $k - 1$ stopni swobody. Ogólna ilość stopni swobody dla tablicy kontyngencji $2 \times k$ wynosi $2k - 1$. Pokażemy teraz sposób wnioskowania na podstawie pozostałych k stopni swobody. Dotyczyć to będzie hipotez nie wykorzystujących sum brzegowych.

Przykład 6.8 (według [Parker])

Badano populację muszek *Drosophila* pochodzącą z skrzyżowania muszek heterozygotycznych ze względu na recesywny gen barwy oczu. Populację podzielono na pięć próbek, każdą zajął się inny student. Chcemy odpowiedzieć na pytanie: czy można uważać, że proporcja pomiędzy fenotypem dzikim i fenotypem mutantów jest równa 3:1, biorąc pod uwagę wszystkie wyniki uzyskane od pięciu różnych osób. Otrzymane dane przedstawia tabela 6.14.

Mijałoby się z sensem stosowanie jakiegokolwiek testu, który „brałby pod uwagę" sumy marginalne dla próbek, gdyż są one całkowicie dowolne. Wobec tego z $2k - 1 = 9$ stopni swobody odpadają $k - 1 = 4$ stopnie swobody, związane z sumami marginalnymi i pozostaje do dyspozycji $k = 5$ stopni swobody.

Obliczymy obecnie wielkość $\chi^2_{3:1}$ dla weryfikacji zgodności obserwowanych sum populacyjnych z sumami populacyjnymi oczekiwanymi (prawy margines tabeli), wynikającymi z hipotezy, że stosunek fenotypu dzikiego i fenotypu mutantów wynosi 3:1. Do obliczeń stosujemy klasyczny wzór:

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

Tabela 6.14

Wyniki badań nad krzyżowaniem muszek Drosophila *dokonanych przez pięciu studentów*

		Próbka					(Razem)
		1	2	3	4	5	
Typ dziki	*O*	60	75	81	70	54	340
	E	64,5	81,0	74,25	84,0	56,25	360
	O – *E*	–4,5	–6,0	+6,25	–14,0	–2,25	–20
Mutanty	*O*	26	33	18	42	21	140
	E	21,5	27,0	24,75	28,0	18,75	120
	O – *E*	+4,5	+6,0	–6,25	+14,0	+2,25	+20
(Razem)		86	108	99	112	75	480

Otrzymujemy:

$$\chi^2_{3:1} = \frac{(-20)^2}{360} + \frac{20^2}{120} = 4{,}444$$

Uwzględniając, że $\chi^2_{3:1}$ ma rozkład χ^2 o jednym stopniu swobody i że

$$_{0{,}05}\chi^2_{(1)} = 3{,}84$$

widzimy,że

$$\chi^2_{3:1} > {}_{0{,}05}\chi^2_{(1)}$$

Odchylenie od stosunku 3:1 wydaje się więc być istotne. Obliczymy teraz poszczególne χ^2 dla testowania zgodności częstości obserwowanych w każdej próbce ze stosunkiem 3:1. Każda z takich wartości ma rozkład χ^2 o jednym stopniu swobody. Następnie zsumujemy otrzymane liczby w celu uzyskania całkowitego χ^2 przy 5 stopniach swobody. Mamy więc:

$$\chi^2_{próbka} = \frac{(-4{,}5)^2}{64{,}5} + \frac{(4{,}5)^2}{21{,}5} = 1{,}256$$

i podobnie:

$$\chi^2_{próbka_2} = 1{,}778$$

$$\chi^2_{próbka_3} = 2{,}455$$

$$\chi^2_{próbka_4} = 9{,}333$$

$$\chi^2_{próbka_5} = 0{,}360$$

oraz

$$\chi^2_{całk} = \sum_{i=1}^{5} \chi_{próbka_i} = 15{,}182$$

Drogą odejmowania znajdujemyd χ^2 dla niejednorodności wyników w poszczególnych próbach

$$\chi^2_{niejedn.} = \chi^2_{całk.} - \chi^2_{3:1} = 10{,}738$$

$\chi^2_{niejedn.}$ ma rozkład χ^2 o 5 – 1 = 4 stopniach swobody. Ponieważ

$$\chi^2_{niejedn.} > {}_{0{,}05}\ \chi^2_{(4)} = 9{,}49$$

więc wnioskujemy, że wyniki poszczególnych próbek są istotnie niejednorodne. Wobec tego wartość rezultatu uzyskanego na podstawie analizy sum wyników wszystkich próbek, a dotyczącego istotności odchylenia od stosunku 3:1 w populacji staje pod znakiem zapytania, gdyż nie można traktować wszystkich próbek jako pochodzących z jednego eksperymentu. Przeglądając wartości χ^2 dla poszczególnych próbek widać, że χ^2 dla próbki czwartej jest bardzo wysokie, istotne na poziomie 0,01

$$\chi^2_{próbka_4} > {}_{0{,}01}\ \chi_{(1)} = 6{,}63$$

natomiast żadna z pozostałych próbek nie wykazuje znamiennego odchylenia od oczekiwanego stosunku 3:1. Czy wolno więc odrzucić wyniki z próbki czwartej, które wnoszą największy wkład w niejednorodność? Można tak zrobić tylko wówczas, gdy mamy inne, wyraźne powody sądzić, że dane są mylne. Załóżmy, że tak jest rzeczywiście. Po wykonaniu badań okazało się, że student opracowujący próbkę czwartą wykazywał upośledzenie w widzeniu barwnym. Wobec tego można było zrezygnować z wyników jego badań, usuwając odpowiednie dane, a pozostałe dane (por. tabela 6.15) poddając ponownej analizie.

Tabela 6.15

Wyniki badań nad krzyżowaniem muszek Drosophila *po usunięciu błędnych danych*

	Próbka				(Razem)	
	1	2	3	4		
Typ dziki	60	75	81	54	(O) (E) $(O-E)$	270 276 –6
Mutanty	26	33	18	21	(O) (E) $(O-E)$	98 92 +6
(Razem)	86	108	99	75	(O)	368

Uzyskuje się następujące wyniki:

$$\chi^2_{całk.} = 1{,}256 + 1{,}778 + 2{,}455 + 0{,}360 = 5{,}849$$

$$\chi^2_{3:1} = 0{,}522$$

$$\chi^2_{niejedn.} = 5{,}849 - 0{,}522 = 5{,}327$$

$$\chi^2_{3:1} < {}_{0{,}05}\,\chi^2_{(1)} = 3{,}84$$

$$\chi^2_{niejedn.} < {}_{0{,}05}\,\chi^2_{(3)} = 7{,}81$$

Można więc wnioskować, że odchylenie od stosunku 3:1 nie jest istotne, oraz że przy eksperymencie uzyskano zadowalającą jednorodność wyników.

6.3 Ogólne tablice kontyngencji *r*×*c*

6.3.1 Test niezależności χ^2

Będziemy rozpatrywać analizę tablic kontyngencji o wymiarach $r \times c$, gdzie r jest liczbą wierszy, a c liczbą kolumn tablicy oraz zarówno r jak i c są większe od dwóch. Test χ^2 dla takich tablic służy do wykrywania związku pomiędzy dwoma klasyfikacjami

typu jakościowego, z których każda zakłada podział na więcej niż dwie kategorie. Obliczona statystyka χ^2 wyraża się znanym wzorem

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

(sumowanie przebiega po wszystkich polach tabeli) i ma rozkład χ^2 o $(r-1)(c-1)$ stopniach swobody. Liczebności oczekiwane E wyznacza się w oparciu o sumy marginalne jako:

$$E = \frac{\text{suma\_wiersza} \cdot \text{suma\_kolumny}}{N}$$

gdzie N jest całkowitą liczebnością obserwacji.

Warunki korzystania z testu χ^2 dla tablic $r \times c$ nie są tak ostre, jak to było w przypadku tablic 2×2. Można więc korzystać z tego testu pod warunkiem, że tylko nieliczne częstości oczekiwane są mniejsze niż 5 (jeden wynik na 5 lub więcej pól w tabeli, względnie 2 wyniki na 10 lub więcej pól) i w wypadku, gdy żadna z częstości oczekiwanych nie jest mniejsza niż 1. Jeżeli te warunki nie są spełnione można połączyć te wiersze lub kolumny, w których występują małe częstości oczekiwane i do takiej zagregowanej tabeli stosować test χ^2.

6.3.2 Miary siły związku

W punkcie 6.1.3 omówiono różnicę pomiędzy stwierdzeniem istotności lub nieistotności związku a oceną siły związku. Tamże przedstawiono kilka wzorów umożliwiających liczbową ocenę siły związku. Mierniki tamte są dogodne do szacowania siły związku dla tablic 2×2, nie nadają się jednakże dla tablic kontyngencji o większej wymiarowości. Określony wzorem (6.10) miernik Φ^2 w ogólnym przypadku tablicy $r \times c$ może znacznie przekroczyć jedność. Dlatego wprowadzono inne mierniki oparte na χ^2, które przyjmują wartości z przedziału od 0 do 1. Jednym z nich jest miernik T Czuprowa określany przez zależność:

$$T^2 = \frac{\chi^2}{N\sqrt{(r-1)(c-1)}} \tag{6.24}$$

Miernik ten może przyjąć wartość maksymalną równą 1 tylko w przypadku tablicy kwadratowej ($r = c$). W każdym innym przypadku wartość T musi być mniejsza od jedności. Wady tej nie posiada miernik V Cramera określany wzorem:

$$V^2 = \frac{\chi^2}{N \cdot \min(r-1, c-1)} \tag{6.25}$$

Jeżeli ilość wierszy jest równa ilości kolumn, obydwa mierniki przyjmują identyczne wartości. Innym miernikiem siły związku jest miernik C Pearsona:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}} \tag{6.26}$$

Miernik ten, podobnie jak poprzednie, przyjmuje wartość zero dla niezależnych klasyfikacji, ale jego wartość maksymalna zależy od ilości wierszy i kolumn. Dlatego też jest on trudny do interpretacji, chyba że przeprowadzimy standaryzację dzieląc jego wartość przez maksymalną wartość możliwą dla danej liczby wierszy i kolumn. Taki dzielnik dla tablic 2×2 wynosi $\sqrt{2}/2$, a przekształcony wzór na wartość standaryzowaną miernika Pearsona dla tablic 2×2 był prezentowany w punkcie 6.1.3 jako (6.12).

Przykład 6.9 (według [Armitage])

Tabela 6.16 przedstawia wyniki badań zależności między grupą krwi a chorobami żołądka. Z badań wyłączono niewielką część osób z grupą krwi AB. Z chorób uwzględniono chorobę wrzodową i raka żołądka. Poza tym określono grupę krwi dla grupy kontrolnej osób zdrowych.W wyniku obliczeń otrzymujemy:

$$\chi^2 = 40{,}54$$

$$\chi^2 > {}_{0{,}001}\chi^2_{(4)} = 18{,}47$$

$$V^2 = T^2 = 0{,}002312$$

$$V = T = 0{,}0481$$

$$C = 0{,}0678$$

Porównując obliczoną wartość statystyki χ^2 z wartościami krytycznymi rozkładu χ^2 widać, że zależność między grupą krwi a chorobami żołądka jest wysoce istotna (na poziomie istotności $\alpha = 0{,}001$). Jednak miary siły związku mają wartości niewielkie, co spowodowane jest tym, że związek nie jest bardzo mocny. Na wysoką wartość χ^2 w tym przypadku znaczny wpływ miała duża liczebność próby (por. punkt 6.1.3). Dalszą analizę danych z tabeli 6.16 przedstawimy w następnym punkcie.

Tabela 6.16

Częstość grup krwi AB0 u pacjentów z chorobami żołądka i w grupie kontrolnej

			Choroba wrzodowa *W*	Rak żołądka *R*	Grupa kontrolna *K*	(Razem)
Grupy krwi	0	*O* *E* *O* – *E*	983 872,39 +110,61	383 428,91 –45,91	2892 2956,70 –64,70	4258
	A	*O* *E* *O* – *E*	679 762,16 –83,16	416 374,72 +41,28	2625 2583,12 +41,88	3720
	B	*O* *E* *O* – *E*	134 161,44 –27,44	84 79,38 +4,62	570 547,18 +22,82	788
(Razem)			1796	883	6087	8766

6.3.3 Wyodrębnianie składników χ^2

Jeżeli badanie tablicy wielopolowej $r \times c$ wykaże istotną zależność między dwoma klasyfikacjami, to często zachodzi potrzeba stwierdzenia, czy zależności te utrzymują się w pewnych fragmentach całej tablicy. Można więc próbować znaleźć podział całkowitego χ^2 na składniki odpowiadające pewnym klasyfikacjom „zagregowanym", dającym po „nałożeniu się" na siebie klasyfikacje pierwotne. W ten sposób istnieje możliwość wykrycia „miejsca" w tabeli, w którym zależność wykryta przez badanie całkowitego χ^2 „skoncentrowała" się. Przykład takiego postępowania pokażemy poniżej. Tutaj należy jeszcze zaznaczyć, że jeżeli do obliczenia składników ogólnego χ^2 zastosujemy liczebności oczekiwane, obliczone dla całej tablicy, to możemy poszczególne składniki dodawać do siebie, a ich suma da dokładnie całkowitą wartość χ^2. Jeżeli jednak liczebności oczekiwane będą obliczone oddzielnie dla poszczególnych „tablic składowych", poszczególne składniki χ^2 nie dadzą w sumie ogólnego χ^2. Rozbieżność ta nie będzie jednak miała większego praktycznego znaczenia.

Przykład 6.9 (kontynuacja z punktu 6.3.2)

Dane zawarte w tabeli 6.16 mogą sugerować istnienie niewielkiej różnicy między pacjentami z rakiem żołądka a grupą kontrolną oraz większą częstość grupy krwi 0 wśród pacjentów z chorobą wrzodową. Dla zbadania takiej sugestii dokonujemy dezagregacji ogólnej tablicy 3×3 na trzy „tablice składowe" — jak to pokazuje tabela 6.18 — i dla każdej obliczamy wartość χ^2. Przykładowo przytoczymy wyniki obliczeń dla wariantu (*b*) — por. także tabela 6.17.

Tabela 6.17

Porównanie częstości występowania grupy krwi 0 wśród pacjentów z chorobą wrzodową z pozostałymi przypadkami — na podstawie danych z tabeli 6.16

			Choroba wrzodowa *W*	Rak żołądka + grupa kontrolna *R + K*	(Razem)
Grupy krwi	0	*O* *E* *O – E*	983 872,39 +110,61	3275 3385,61 –110,61	4258
	A + B	*O* *E* *O – E*	813 923,61 –110,61	3695 3584,39 +110,61	4508
(Razem)			1796	6970	8766

Tabela 6.18

Schemat podziału tablicy kontyngencji 3 × 3 z tabeli 6.16 dla dokonania szczegółowych porównań

(a)

	W	*R*	*K*
0	x	x	x
A	x	x	x
B	x	x	x

(b)

	W	*R*	*K*
0	x	x	x
A	x	x	x
B	x	x	x

(c)

	W	*R*	*K*
0	x	x	x
A	x	x	x
B	x	x	x

Mamy:

$$\chi^2_{(b)} = 34{,}29$$

$$\chi^2_{(b)} > {}_{0{,}001}\,\chi^2_{(1)} = 10{,}83$$

$$V^2 = T^2 = 0{,}003912$$

$$V = T = 0{,}0625$$

$$C = 0{,}0624$$

Uzyskaliśmy tutaj równie silne jak poprzednio dla całej tablicy 3 × 3 potwierdzenie istotności związku. Porównanie wartości mierników Czuprowa i Cramera wskazują na pewne powiększenie siły związku w porównaniu z przypadkiem tablicy pełnej. Wartość miernika C Pearsona nie może stanowić podstawy dla porównań siły związku w tabelach kontyngencji o różnej wymiarowości (por. punkt 6.1.3).

Tabela 6.19

Zestawienie wartości składników χ^2 dla trzech porównań dokonanych zgodnie z tabelą 6.18 dla danych z tabeli 6.16

Układ	Porównanie w wierszach	Porównanie w kolumnach	Składnik χ^2	Liczba stopni swobody	Wartość krytyczna χ^2 dla $\alpha = 0{,}05$	Istotność
(a)	0 z A z B	R z K	5,61	2	5,99	
(b)	0 z $(A+B)$	W z $(R+K)$	34,29	1	3,84	istotne
(c)	A z B	W z $(R+K)$	0,68	1	3,84	
			40,61	4		

Zestawienie wyników porównania wszystkich trzech tablic składowych przedstawia tabela 6.19. Widać wyraźne potwierdzenie istotności zależności badanej w układzie (b) — tzn. większej częstości grupy krwi 0 wśród chorych na wrzody żołądka w porównaniu z pozostałymi możliwymi przypadkami. Dla układu (a) uzyskaliśmy wartość χ^2 nieistotną wprawdzie na poziomie 5%, ale niewiele odbiegającą od wartości

krytycznej, co świadczy o tym, że nasze początkowe sugestie nie były zupełnie nieuzasadnione.

Suma składników χ^2 nieco odbiega od wartości obliczonej dla całej tablicy 3×3, co zostało wyjaśnione w tekście powyżej. Rozbieżność ta jest do pominięcia.

7. ANALIZA WARIANCJI

7.1 Analiza wariancji w klasyfikacji pojedynczej

W podrozdziale 5.3. omówiono sposób porównywania dwóch wartości średnich. Często jednakże zachodzi potrzeba porównania większej ilości średnich i udzielenia odpowiedzi na pytanie: czy średnie te różnią się istotnie od siebie, czy też nie. Do odpowiedzi na tego typu pytanie wykorzystujemy zespół metod statystycznych zwanych analizą wariancji. Proste porównywanie kilku średnich nie jest jedynym zastosowaniem analizy wariancji. Metoda ta, w ogólnym przypadku, pozwala na sprawdzenie, czy pewne czynniki wywierają wpływ, a jeśli tak, to jaki wielki, na kształtowanie się średnich wartości badanych cech mierzalnych. Testy analizy wariancji są podstawowym narzędziem statystyki eksperymentalnej, czyli statystycznej metody planowania i oceny wyników eksperymentów naukowych. Niektóre z tych testów będą szczegółowo przedstawione w dalszych częściach niniejszego rozdziału. Obecnie zajmiemy się sposobami porównywania kilku średnich.

7.1.1 Porównywanie kilku średnich

Będziemy zakładać, że rozpatrujemy k grup obserwacji o charakterze ilościowym. W każdej i-tej grupie dysponujemy próbką zawierającą n_i obserwacji. Zakładamy dalej, że obserwacje w każdej grupie mają rozkład normalny lub zbliżony do normalnego, zaś wariancje we wszystkich grupach są równe i wynoszą σ^2, choć σ^2 nie musi być znane. Niech y_{ij} oznacza j-tą obserwację w i-tej grupie. Dalsze oznaczenia pokazano w tabeli 7.1. Jeżeli przez μ_i oznaczymy „rzeczywistą" średnią w i-tej grupie to można przyjąć model, w którym

$$y_{ij} = \mu_i + \varepsilon_{ij} \tag{7.1}$$

gdzie ε_{ij} jest składnikiem losowym o rozkładzie normalnym z zerową wartością średnią i wariancją σ^2. Na podstawie znajomości y_{ij} mamy zweryfikować hipotezę zerową mówiącą, że wszystkie średnie μ_i w grupach są równe

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$$

Tabela 7.1

Oznaczenia stosowane w analizie wariancji w klasyfikacji pojedynczej

Grupa	1	2	... i ...	k	Wszystkie grupy łącznie
liczba obserwacji	n_1	n_2	n_i	n_k	$N=\sum_{i=1}^{k} n_i$
średnia zmiennej y	$\bar{y}_1$	$\bar{y}_2$	$\bar{y}_i$	$\bar{y}_k$	$\bar{y}=\frac{T}{N}$
suma zmiennej y	T_1	T_2	T_i	T_k	$T=\sum_{i=1}^{k} T_i$
suma y^2	S_1	S_2	S_i	S_k	$S=\sum_{i=1}^{k} S_i$
	$T_i=\sum_{j=1}^{n_i} y_{ij}$			$S_i=\sum_{j=1}^{n_i} y_{ij}^2$	

wobec hipotezy alternatywnej H_1 : nie wszystkie średnie grupowe są równe.

Zależność (7.1) przedstawiającą „model" koncepcyjny można przedstawić jako:

$$y_{ij}=\mu+\alpha_i+\varepsilon_{ij} \tag{7.2}$$

gdzie μ jest wielkością stałą niezależną od grupy i tak dobraną, aby

$$\sum_{i=1}^{k} \alpha_i = 0$$

zaś α_i odpowiadają za systematyczne różnice między grupami. Wykorzystując model (7.2) i przechodząc do danych otrzymanych z próby można zauważyć że odchylenie każdej obserwacji od „ogólnej średniej" $\bar{y}$ może być rozdzielone na dwie części: odchylenie obserwacji y_{ij} od średniej grupowej $\bar{y}_i$ oraz odchylenie średniej grupowej od średniej ogólnej $\bar{y}$.

$$y_{ij}-\bar{y}=(y_{ij}-\bar{y}_i)+(\bar{y}_i-\bar{y})^2 \tag{7.3}$$

Można udowodnić, że podnosząc (7.3) do kwadratu i sumując po wszystkich obserwacjach uzyskamy zależność

$$\sum_{i,j}(y_{ij}-\bar{y})^2=\sum_{i,j}(y_{ij}-\bar{y}_i)^2+\sum_{i,j}(\bar{y}_i-\bar{y})^2 \tag{7.4}$$

Oznacza to, że całkowita suma kwadratów odchyleń od średniej może być rozbita na sumę kwadratów odchyleń obserwacji od średnich grupowych oraz sumę kwadratów odchyleń średnich grupowych od średniej ogólnej. Pierwsza z sum składowych jest miarą zmienności wewnątrz grup a druga miarą różnic pomiędzy grupami. Oznaczmy całkowitą sumę kwadratów jako *SK*, sumę kwadratów wewnątrz grup jako *SKWG* oraz sumę kwadratów między grupami jako *SKMG*. Praktyczne wzory do obliczania poszczególnych sum kwadratów są następujące:

$$SK=\sum_{i,j}(y_{ij}-\bar{y})^2=S-\frac{T^2}{N} \tag{7.5}$$

$$SKWG=\sum_{i,j}(y_{ij}-\bar{y}_i)^2=S-\sum_{i}\frac{T_i^2}{n_i} \tag{7.6}$$

$$SKMG=\sum_{i,j}(\bar{y}_i-\bar{y})^2=\sum_{i}\frac{T_i^2}{n_i}-\frac{T^2}{N} \tag{7.7}$$

Przypomnijmy, że założyliśmy na wstępie, iż wariancja σ^2 jest taka sama dla wszystkich grup. Jeżeli hipoteza zerowa jest prawdziwa to istnieją trzy nieobciążone estymatory tej samej wartości σ^2 (oparte na trzech powyższych sumach kwadratów). Są to: tzw. całkowity średni kwadrat s_T^2

$$s_T^2=\frac{SK}{N-1} \tag{7.8}$$

średni kwadrat wewnątrz grup s_W^2

$$s_W^2=\frac{SKWG}{N-k} \tag{7.9}$$

oraz średni kwadrat między grupami s_M^2

$$s_M^2 = \frac{SKMG}{k-1} \tag{7.10}$$

Jeżeli hipoteza zerowa nie jest prawdziwa, to wtedy średni kwadrat wewnątrzgrupowy jest nadal nieobciążonym estymatorem σ^2, natomiast średni kwadrat między grupami opierający się na zmienności między średnimi grupowymi będzie wzrastał. Jeżeli więc s_M^2 znacznie przewyższa s_W^2 hipotezę zerową należy odrzucić. Test opiera się na wartości ilorazu

$$F = \frac{s_M^2}{s_W^2} \tag{7.11}$$

który ma rozkład F Snedecora o $k-1$ i $N-k$ stopniach swobody. Już wartości stosunku F większe od jedności wskazują na niezgodność z hipotezą zerową. Jednakże hipotezę zerową możemy odrzucić i automatycznie przyjąć hipotezę alternatywną dopiero wtedy, gdy wartość F będzie dostatecznie duża i przekroczy wartość krytyczną rozkładu F przy zadanym poziomie istotności i o podanej wyżej liczbie stopni swobody. Gdy

$$F < {}_{\alpha}F_{(N-k)}^{(k-1)}$$

nie ma podstaw do odrzucenia hipotezy zerowej.

Wyniki analizy wariancji zapisujemy zwykle według schematu zwanego tablicą analizy wariancji. Dla omawianej analizy wariancji w klasyfikacji pojedynczej tablica taka ma postać pokazaną w tabeli 7.2.

Tabela 7.2

Tabela analizy wariancji w klasyfikacji pojedynczej

Źródło zmienności	Suma kwadratów	Liczba stopni swobody	Średni kwadrat	Stosunek wariancji
Między grupami	$SKMG$	$k-1$	$s_M^2 = \frac{SKMG}{k-1}$	$F = \frac{s_M^2}{s_W^2}$
Wewnątrz grup	$SKWG$	$N-k$	$s_W^2 = \frac{SKWG}{N-k}$	
Całkowita	SK	N – 1		

Przykład 7.1 (według [Armitage])

Czterema różnymi metodami mierzono czas krzepnięcia osocza. Osocze pobierano od dziesięciu pacjentów i poddawano czterem testom. Wyniki przedstawiono w tabeli 7.3.

Tabela 7.3

Czas krzepnięcia osocza (w min.) mierzony czterema metodami

Metoda	1	2	3	4
	9,1	10,0	10,0	10,9
	8,9	10,2	9,9	11,1
	8,4	9,8	9,8	12,2
	12,8	11,6	12,9	14,4
	8,7	9,5	11,2	9,8
	9,2	9,2	9,9	12,0
	7,6	8,6	8,5	8,5
	8,6	10,3	9,8	10,9
	8,9	9,4	9,2	10,4
	7,9	8,5	8,2	10,0
T_i	90,1	97,1	99,4	110,2
T_i^2	8118,01	9428,41	9880,36	12144,04
$\bar{y}_i$	9,01	9,71	9,94	11,02

Należy porównać średnie czasy krzepnięcia osocza uzyskane w wyniku pomiaru każdą z metod. Na podstawie danych z tabeli 7.3 obliczono:

$$N = 40$$

$$T = 396{,}8$$

$$S = 4021{,}84$$

$$\frac{T^2}{N} = 3936{,}256$$

$$\sum \frac{T_i^2}{n_i} = 3957{,}082$$

Tablicę analizy wariancji dla rozważanego przykładu przedstawiono w tabeli 7.4.

Tablica 7.4

Tablica analizy wariancji dla danych z tabeli 7.3

Źródło zmienności	Suma kwadratów	Liczba stopni swobody	Średni kwadrat	Stosunek wariancji	Istotność
Między metodami	20,826	3	6,924	3,86	$P < 0{,}025$
Wewnątrz metod	64,758	36	1,799		
Całkowita	85,584	39			

Ponieważ

$$F > {}_{0{,}025}F^{(3)}_{(36)}$$

więc hipotezę zerową o równości średnich uzyskanych w wyniku pomiaru czterema metodami należy odrzucić. Dalszą analizę tych danych przeprowadzimy w następnym punkcie.

7.1.2 Wyodrębnianie kontrastów liniowych

Jeżeli analiza wariancji nie wskaże istotności różnic między średnimi, nie prowadzimy już dalszych testów. Gdy natomiast procedura analizy wariancji daje istotny wynik testu F zachodzi potrzeba dokładniejszego zbadania różnic między poszczególnymi średnimi grupowymi. Chcąc porównać np. średnie $\bar{y}_g$ i $\bar{y}_h$ można skorzystać z testu t

$$t = \frac{\bar{y}_g - \bar{y}_h}{s_w \sqrt{\frac{1}{n_g} - \frac{1}{n_h}}} \tag{7.12}$$

o $N - k$ stopniach swobody dla oceny istotności różnicy między tymi średnimi. Wzór ten różni się od podanego w podrozdziale 5.3 innym sposobem określenia błędu standardowego różnicy średnich. Wykorzystuje się tutaj średni kwadrat wewnątrz grup s_W^2 o $N - k$ stopniach swobody.

Jeżeli wszystkie grupy są równoliczne ($n_1 = n_2 = \ldots = n_k = n$) to wygodnie jest obliczyć najmniejszą istotną różnicę między średnimi przy określonym poziomie istotności α. Różnicę tę oznaczoną tutaj jako D można wyrazić wzorem:

$$D = \alpha t_{(N-k)} \cdot s_W \sqrt{\frac{2}{n}} \tag{7.13}$$

Istotne będą wszystkie te różnice między poszczególnymi średnimi, które przekraczają wartość D.

Gdy chcemy dokonać bardziej zaawansowanych porównań, np. między grupami średnich korzystamy z tzw. kontrastów liniowych

$$L = \sum_{i=1}^{k} \lambda_i \bar{y}_i \tag{7.14}$$

gdzie

$$\sum_{i=1}^{k} \lambda_i = 0 \tag{7.15}$$

Kontrasty można tworzyć w różny sposób [Armitage]. Tutaj podamy jeden ze sposobów w charakterze przykładu.

Będziemy dalej nadal rozważać sytuację, w której wszystkie grupy są równoliczne, tzn.

$$n_1 = n_2 = \ldots = n_k = n$$

Przypuśćmy, że interesuje nas porównanie średniej $\bar{y}_c$ jednej wybranej grupy (np. kontrolnej) ze średnią pewnego podzbioru q innych grup (podzbioru wybranego spośród wszystkich pozostałych $k - 1$ grup). Chcąc więc rozważać różnicę

$$\bar{y}_c - \frac{1}{q} \sum_{i=1}^{q} \bar{y}_i$$

możemy utworzyć kontrast

$$L = q \cdot \bar{y}_c - \sum_{i=1}^{q} \bar{y}_i \tag{7.16}$$

powstały z pomnożenia badanej różnicy przez q. Współczynniki λ_i w tym kontraście są równe -1 dla q średnich grupowych przeciwstawianych średniej $\bar{y}_c$ i 0 dla pozostałych, nie biorących udziału w porównywaniu, średnich.

Standardowy błąd kontrastu L szacuje się jako

$$s_{(L)} = s_W \sqrt{\frac{\sum_i \lambda_i^2}{n}} \tag{7.17}$$

Badanie istotności kontrastu polega na rozbiciu uzyskanej w wyniku analizy wariancji sumy kwadratów między grupami na dwa składniki: sumę kwadratów względem kontrastu L oraz sumę kwadratów względem wszystkich innych kontrastów. Składniki te oblicza się z zależności:

$$SKL = \frac{L^2}{\frac{1}{n}\sum \lambda_i^2} \tag{7.18}$$

$$SKP = SKMG - SKL \tag{7.19}$$

gdzie:

SKL — suma kwadratów względem kontrastu L

SKP — suma kwadratów względem innych kontrastów.

Następnie tworzy się odpowiednie średnie kwadraty

$$s_1^2 = \frac{SKL}{1} = SKL \tag{7.20}$$

$$s_2^2 = \frac{SKP}{k-2} \tag{7.21}$$

dzieląc sumy kwadratów przez liczby stopni swobody. Dalej tworzy się dwa stosunki oszacowań wariancji F_1 i F_2 dzieląc średnie kwadraty (7.20) i (7.21) przez średni kwadrat wewnątrz grup s_W^2. Dla F_1 i F_2 przeprowadza się dwa odrębne testy istotności, przy czym F_1 przy 1 i $N - k$ stopniach swobody, a F_2 przy $k - 2$ i $N - k$ stopniach swobody. Jeżeli ogólny test F (7.11) dał wynik istotny, to należy się spodziewać, że co najmniej jeden z testów F_1 lub F_2 powinien dać także wynik istotny pozwalając potwierdzić lub odrzucić nasze przypuszczenia stojące u podstaw konstrukcji kontrastu liniowego L. Tabela 7.5 pokazuje tablicę analizy wariancji dla przypadku badania kontrastu liniowego L.

Tabela 7.5

Tablica analizy wariancji dla badania istotności kontrastu liniowego L

Źródło zmienności	Suma kwadratów	Liczba stopni swobody	Średni kwadrat	Stosunek wariancji
Względem kontrastu L	SKL	1	$s_1^2 = SKL$	$F_1 = \frac{s_1^2}{s_W^2}$
Względem innych kontrastów	SKP	$k - 2$	$s_2^2 = \frac{SKP}{k-2}$	$F_2 = \frac{s_2^2}{s_W^2}$
Wewnątrz grup	$SKWG$	$N - k = k(n - 1)$	$s_W^2 = \frac{SKWG}{k(n-1)}$	
Całkowita	SK	$N - 1 = nk - 1$		

Przykład 7.1 (kontynuacja z poprzedniego punktu)

W danych dotyczących czasu krzepnięcia osocza z tabeli 7.2 zaobserwowano znaczne odchylenie średniego czasu krzepnięcia dla metody 4 w porównaniu z metodami 2 i 3, jak również mniejsze odchylenie średniego czasu krzepnięcia uzyskanego metodą 1 w stosunku do metod 2 i 3. Postanowiono więc porównać, wykorzystując kontrasty liniowe, każdą ze średnich dla metod „skrajnych" (1 i 4) ze średnimi dla metod „środkowych" (2 i 3) oraz każdą ze średnich „skrajnych" z pozostałymi średnimi. Utworzono 4 kontrasty liniowe

$$L_{(1)} = 2\overline{y}_1 - \overline{y}_2 - \overline{y}_3$$
$$L_{(2)} = -\overline{y}_2 - \overline{y}_3 + 2\overline{y}_4$$
$$L_{(3)} = -\overline{y}_1 - \overline{y}_2 - \overline{y}_3 + 3\overline{y}_4$$
$$L_{(4)} = 3\overline{y}_1 - \overline{y}_2 - \overline{y}_3 - \overline{y}_4$$

Ich wartości są następujące

$L_{(1)} = -1{,}63$ $\qquad$ $L_{(3)} = 4{,}4$

$L_{(2)} = 2{,}39$ $\qquad$ $L_{(4)} = -3{,}64$

zaś odpowiednie tablice analizy wariancji przedstawia tabela 7.6. Widać, że istotne (zgodnie z wykonanym testem F) są trzy ostatnie kontrasty, zaś istotność ta jest najsilniejsza dla kontrastu $L_{(3)}$ ($\alpha = 0{,}01$), czyli dla przeciwstawienia średniej $\bar{y}_4$ wszystkim pozostałym średnim.

Tabela 7.6

Tablice analizy wariancji dla badania czterech kontrastów dla danych z tabeli 7.3

Sumy kwadratów	Liczby stopni swobody	Średnie kwadraty	Stosunki wariancji	Istotność
KONTRAST $L_{(1)}$				
$SKL = 4{,}428$ $SKP = 16{,}398$ $SKWG =$ $= 64{,}758$	1 2 36	$s_1^2 = 4{,}428$ $s_2^2 = 8{,}199$ $s_W^2 = 1{,}799$	$F_1 = 2{,}46$ $F_2 = 4{,}56$	nieistotne $P < 0{,}025$
KONTRAST $L_{(2)}$				
$SKL = 9{,}520$ $SKP = 11{,}306$ $SKWG =$ $= 64{,}758$	1 2 36	$s_1^2 = 9{,}520$ $s_2^2 = 5{,}653$ $s_W^2 = 1{,}799$	$F_1 = 5{,}29$ $F_2 = 3{,}14$	$P < 0{,}05$ nieistotne $(P > 0{,}05)$
KONTRAST $L_{(3)}$				
$SKL = 16{,}133$ $SKP = 4{,}693$ $SKWG =$ $= 64{,}758$	1 2 36	$s_1^2 = 16{,}133$ $s_2^2 = 2{,}346$ $s_W^2 = 1{,}799$	$F_1 = 8{,}96$ $F_2 = 1{,}30$	$P < 0{,}01$ nieistotne
KONTRAST $L_{(4)}$				
$SKL = 11{,}041$ $SKP = 9{,}785$ $SKWG =$ $= 64{,}758$	1 2 36	$s_1^2 = 11{,}041$ $s_2^2 = 4{,}892$ $s_W^2 = 1{,}799$	$F_1 = 6{,}14$ $F_2 = 2{,}72$	$P < 0{,}025$ nieistotne

Jeżeli dla pewnego zestawu danych, po dokonaniu pierwotnej analizy wariancji, wykonujemy wiele porównań a posteriori badając wiele różnych kontrastów, to znacznie wzrasta prawdopodobieństwo popełnienia błędu polegającego na przypadkowym uznaniu

pewnego lub pewnych kontrastów za istotne. Dlatego w przypadku dokonywania wielokrotnych porównań, już po zapoznaniu się z danymi (czyli a posteriori), zaleca się stosowanie bardziej ostrożnego testu Scheffe'go, który uznaje liniowy kontrast L za istotny na poziomie α, gdy jest spełnione

$$\left|\frac{L}{s_{(L)}}\right| \geq \sqrt{(k-1)\cdot \alpha F_{(N-k)}^{(k-1)}} \tag{7.22}$$

Przykład 7.1 (dokończenie)

Jak wynika z powyższych stwierdzeń nasze 4 kontrasty powinniśmy testować raczej testem Scheffe'go, niż klasycznym testem F. Prawa część nierówności (7.22) w naszym przypadku wynosi:

$$\sqrt{(4-1)\cdot {}_{0,05}F_{(36)}^{(3)}} = 2{,}934$$

zaś wartości lewych stron (7.22) dla poszczególnych kontrastów przedstawione są w tabeli 7.7. Uzyskaliśmy wiarygodne potwierdzenie istotności trzeciego kontrastu, natomiast istotność kontrastów drugiego i czwartego — wskazywana przez wyodrębnianie odpowiedniego składnika międzygrupowej sumy kwadratów — nie potwierdziła się.

Tabela 7.7

Wyniki testu Scheffe'go dla czterech kontrastów liniowych dotyczących danych z tabeli 7.3

Kontrast	$\left\|\frac{L}{s_{(L)}}\right\|$	Istotność
$L_{(1)}$	1,569	nieistotne
$L_{(2)}$	2,300	nieistotne
$L_{(3)}$	2,995	istotne dla $\alpha = 0{,}05$
$L_{(4)}$	2,477	nieistotne

Metodę Scheffe'go oraz inne „ostrożne" metody wielokrotnego porównywania (np. metodę rozstępu studentyzowanego, często opisywaną w literaturze, np. [Armitage], [Parker]) stosuje się wtedy, gdy dokonujemy wielu porównań dla wykrycia źródeł „zakłóceń", które zaobserwowaliśmy analizując wyniki doświadczenia lub badania (po ich otrzymaniu) i porównując je z np. pierwotnymi hipotezami o charakterze przyczynowo-skutkowym, modelem zjawiska, itd.

Jeżeli porównywanie średnich lub badanie kontrastów wynika z przyjętego a priori pierwotnego planu doświadczenia lub badania, bardziej właściwe jest stosowanie testu t (7.12), testu najmniejszej istotnej różnicy (7.13) lub testu F połączonego z rozbijaniem sumy kwadratów względem kontrastu (kontrastów).

7.1.3 Test jednorodności wielu wariancji (test Bartletta)

Z punktu 7.1.1 pamiętamy, że jednym z założeń w analizie wariancji jest równość wariancji we wszystkich k grupach obserwacji. Do sprawdzenia tego założenia można użyć testu jednorodności wariancji, zwanego — od nazwiska autora — testem Bartletta.

Test ten stosujemy do porównania wariancji w k grupach obserwacji wybranych losowo ze zbiorowości o rozkładzie normalnym. Oznaczymy liczebność i-tej grupy obserwacji przez n_i. Niech dalej j-ta obserwacja w i-tej grupie oznaczona będzie przez y_{ij}, zaś suma liczebności wszystkich grup — jako N. Testem Barletta sprawdzamy hipotezę zerową o równości wariancji we wszystkich zbiorowościach, z których pochodzą poszczególne grupy obserwacji, wobec hipotezy alternatywnej mówiącej, że nie wszystkie te wariancje są równe. Test polega na obliczeniu (według znanego wzoru) oszacowań wariancji w każdej grupie

$$s_i^2 = \frac{\sum_{i=1}^{n_i} y_{ij}^2 - \frac{\left(\sum_{i=1}^{n_i} y_{ij}\right)^2}{n_i}}{n_i - 1}$$

a następnie wyznaczeniu wielkości $\tilde{s}^2$, M i C:

$$\tilde{s}^2 = \frac{\sum_{i=1}^{k} \left(n_i - 1\right) s_i^2}{N - k} \tag{7.24}$$

$$M = (N - k) \ln \tilde{s}^2 - \sum_{i=1}^{k} \left(n_i - 1\right) \ln s_i^2 \tag{7.25}$$

$$C = 1 + \frac{1}{3\,(k-1)} \left[\sum_{i=1}^{k} \frac{1}{n_i - 1} - \frac{1}{N - k} \right] \tag{7.26}$$

Dalej oblicza się wartość statystyki χ^2

$$\chi^2 = \frac{M}{C} \tag{7.27}$$

Statystyka ta ma w przybliżeniu rozkład χ^2 o $k - 1$ stopniach swobody. Jeżeli obliczona wartość χ^2 jest równa lub przekracza wartość krytyczną $_{\alpha}\chi^2_{(k-1)}$ musimy odrzucić hipotezę zerową o równości wariancji w grupach, w przeciwnym przypadku nie ma podstaw do odrzucenia hipotezy zerowej.

Należy podkreślić, że test Bartletta, jak wszystkie testy dotyczące wariancji, jest wrażliwy na nienormalność rozkładów populacji.

Przykład 7.2

Dla danych z przykładu 7.1 (por. tabela 7.3) należy, stosując test Bartletta, sprawdzić równość wariancji w grupach odpowiadających poszczególnym metodom pomiaru czasu krzepnięcia osocza krwi. Potrzebne pośrednie wyniki obliczeń przedstawia tabela 7.8.

Tabela 7.8

Zestawienie pośrednich wyników obliczeń dla potrzeb testu Bartletta dla danych z tabeli 7.4

Metoda	1	2	3	4
n_i	10	10	10	10
$\sum_j y_{ij}$	90,1	97,1	99,4	110,2
$\sum_j y_{ij}^2$	830,09	950,19	1004,08	1237,04
$\left(\sum_j y_{ij}\right)^2$	8118,01	9428,41	9880,36	12144,04

Uwzględniając je otrzymujemy:

$\tilde{s}^2 = 1{,}799$

$M = 2{,}9007$

$C = 1{,}0462$

$\chi^2 = 2{,}7726$

Ponieważ

$$\chi^2 < {}_{0,05}\chi^2_{(3)} = 7{,}815$$

więc nie ma podstaw do odrzucenia hipotezy o równości wariancji dla poszczególnych metod. Założenie o równości wariancji w grupach było więc zachowane w analizie wariancji przeprowadzonej na danych z tabeli 7.3 w przykładzie 7.1.

7.2. Analiza wariancji w klasyfikacji podwójnej

W podrozdziale 7.1 został omówiony najprostszy przypadek zastosowania analizy wariancji: dla porównania kilku średnich. W przykładzie dotyczącym pomiarów czasu krzepnięcia osocza chciano sprawdzić, czy pewien pojedynczy czynnik wywiera istotny wpływ na kształtowanie się średnich wartości badanej cechy mierzalnej. Tym pojedynczym czynnikiem była zastosowana metoda pomiaru. Przyjęto wówczas model (por.wzór(7.2)), w którym odchylenie obserwacji od średniej ogólnej było w części tłumaczone wpływem tego czynnika, a w pozostałej części zmiennością „losową". Istotą analizy wariancji było rozbicie miary całkowitej zmienności próby (*SK*) na addytywne składniki: składnik wynikający z działania naszego składnika (*SKMG*) i składnik resztowy (*SKWG*), mierzący zmienność losową. Porównanie wariancji wynikającej z wpływu czynnika różnicującego s^2_M z wariancją resztową s^2_W dokonane przy pomocy testu F, dało odpowiedź, czy dany wynik odgrywa istotną rolę w kształtowaniu się średnich wartości badanej cechy.

Gdy zachodzi potrzeba zbadania istotności wpływu dwu różnych czynników na zachowanie się wartości średnich pewnej cechy mierzalnej — stosuje się analizę wariancji w klasyfikacji podwójnej. Przykładem może być rozpatrywanie średniego czasu krzepnięcia krwi w zależności od kilku różnych czasów przechowywania osocza. Schemat postępowania jest bardzo podobny. Różnica zaś polega na rozbiciu całkowitej sumy kwadratów odchyleń od średniej ogólnej na trzy składniki: dwa stanowiące miary wpływu obu czynników oraz składnik resztowy. Następnie porównuje się testami F każdą z dwóch wariancji odpowiadających składnikom związanym z czynnikami różnicującymi z wariancją resztową. Daje to możliwość oceny istotności wpływu każdego czynnika na różnicowanie się wartości średnich.

7.2.1 Schemat addytywny

Klasyfikację obserwacji zmiennej losowej y na r grup związanych z działaniem pierwszego czynnika i c grup związanych z drugim czynnikiem można przedstawić w postaci tablicy o wymiarach $r \times c$, w której wiersze odpowiadają kategoriom pierwszej klasyfikacji, kolumny — drugiej. Przeprowadzimy teraz analizę przypadku, gdy w każdej podgrupie klasyfikacyjnej (w każdym polu tablicy odpowiadającym i-temu wierszowi i j-tej kolumnie

występuje dokładnie jedna obserwacja. Taki układ danych nazywamy blokiem zrandomizowanym. Obecność pojedynczych obserwacji może być spowodowana np. dużymi kosztami eksperymentu; pojedyncze dane mogą być też średnimi z kilku pomiarów. Mamy wobec tego

$$N = r \cdot c$$

obserwacji. Przyjmujemy oznaczenia jak w tabeli 7.9. Przyjmujemy założenie, że każda z $r \cdot c$ podrup klasyfikacyjnych charakteryzuje się rozkładem normalnym ze średnią μ_{ij} i wariancją σ^2, przy czym wariancja jest taka sama we wszystkich podrupach.

Tabela 7.9

Oznaczenia przyjęte w analizie wariancji w klasyfikacji podwójnej przy zastosowaniu modelu addytywnego

kolumny / wiersze	1	2	... j ...	c	suma	Średnia (R_i/c)
1	y_{11}	y_{12}	y_{1j}	y_{1c}	R_1	$y_{1\cdot}$
2	y_{21}	y_{22}	y_{2j}	y_{2c}	R_2	$y_{2\cdot}$
:						
i	y_{i1}	y_{i2}	y_{ij}	y_{ic}	R_i	$y_{i\cdot}$
:						
r	y_{r1}	y_{r2}	y_{rj}	y_{rc}	R_r	$y_{r\cdot}$
suma	C_1	C_2	C_j	C_c	T	
średnia (C_j/r)	$y_{\cdot 1}$	$y_{\cdot 2}$	$y_{\cdot j}$	$y_{\cdot c}$		$\bar{y} = \frac{T}{N}$

$$C_j = \sum_{i=1}^{r} y_{ij} \qquad R_j = \sum_{j=1}^{c} y_{ij}$$

$$\bar{y}_{\cdot j} = \frac{1}{r} C_j \qquad \bar{y}_{i\cdot} = \frac{1}{c} R_i$$

$$N = r \cdot c \qquad T = \sum_i R_i = \sum_j C_j \qquad S = \sum_{i,j} y_{ij}^2$$

Zakłada się dalej addytywny model w postaci:

$$\mu_{ij} = E(y_{ij}) = \mu + \alpha_i + \beta_j \tag{7.28}$$

gdzie α_i i β_j są odpowiednio stałymi odpowiedzialnymi za „efekt" wierszy czyli różnice między średnimi względem pierwszej klasyfikacji, i podobny „efekt kolumn". Dobierając odpowiednie μ można uzyskać:

$$\sum_{i=1}^{r} \alpha_i = 0 \qquad \text{i} \qquad \sum_{j=1}^{c} \beta_j = 0$$

Każda obserwowana wartośc y_{ij} może być na podstawie (7.28) wyrażona zależnością

$$y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \tag{7.29}$$

gdzie ε_{ij} są niezależnymi zmiennymi losowymi o rozkładzie N(0, σ) odpowiedzialnymi za efekt resztowy nie wyjaśniony efektem wierszy i efektem kolumn.

Testuje się hipotezę zerową o jednorodności wszystkich średnich μ_{ij}, tzn:

$$H_0 : \mu_{11} = \mu_{12} = \cdots = \mu_{ij} = \cdots = \mu_{rc}$$

co przy przyjętym modelu (7.28) oznacza, że
— wszystkie α_i odpowiadające efektowi wierszy są równe zero,
— wszystkie β_j odpowiadające efektowi kolumn są równe zero.
Faktycznie więc hipoteza zerowa rozpada się na dwie hipotezy składowe. Również hipotezę alternatywną mówiącą, że nie wszystkie średnie μ_{ij} są równe, można rozdzielić na dwie składowe stanowiące, że:
— istnieje niezerowy efekt wierszy oraz
— istnieje niezerowy efekt kolumn.
Dla zweryfikowania postawionej hipotezy zerowej pokazuje się, że prawdziwą jest zależność

$$\sum_{i,j}\left(y_{ij} - \overline{y}\right)^2 = \sum_{i,j}\left(\overline{y}_{i\cdot} - \overline{y}\right)^2 + \sum_{i,j}\left(\overline{y}_{\cdot j} - \overline{y}\right)^2 + \sum_{i,j}\left(y_{ij} - \overline{y}_{i\cdot} - \overline{y}_{\cdot j} + \overline{y}\right)^2 \tag{7.30}$$

będąca wyrazem możliwości rozbicia sumy kwadratów odchyleń poszczególnych obserwacji od ogólnej średniej, czyli miary całkowitej zmienności próby

$$SK = \sum_{i,j}\left(y_{ij} - \overline{y}\right)^2 = S - \frac{T^2}{N} \tag{7.31}$$

na trzy składniki:

$$SKMW = \sum_{i,j} \left(\bar{y}_{i.} - \bar{\bar{y}} \right)^2 = \frac{\sum_i R_i^2}{c} - \frac{T^2}{N} \tag{7.32}$$

$$SKMK = \sum_{i,j} \left(\bar{y}_{.j} - \bar{\bar{y}} \right)^2 = \frac{\sum_j C_j^2}{r} - \frac{T^2}{N} \tag{7.33}$$

$$SKR = \sum_{i,j} \left(y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{\bar{y}} \right)^2 = SK - SKMW - SKMK \tag{7.34}$$

Składniki te to :

— suma kwadratów między wierszami *SKMW* stanowiąca miarę zmienności wywołanej klasyfikacją pierwszą, (miarę efektu wierszy)

— suma kwadratów między kolumnami *SKMK* stanowiącą miarę zmienności wywołanej klasyfikacją drugą (miarę efektu kolumn)

— resztową suma kwadratów — *SKR* stanowiąca miarę zmienności nie wyjaśnionej efektem wierszy i efektem kolumn.

Sumy kwadratów oblicza się według praktycznych wzorów (7.31)...(7.34), zaś odpowiadające im oszacowania wariancji dzieląc sumy kwadratów przez liczby stopni swobody:

$$s_R^2 = \frac{SKMW}{r-1} \tag{7.35}$$

$$s_C^2 = \frac{SKMW}{c-1} \tag{7.36}$$

$$s_0^2 = \frac{SKMW}{(r-1)(c-1)} \tag{7.37}$$

Przy założeniu prawdziwości hipotezy zerowej oszacowania s_R^2, s_C^2, s_0^2 są nieobciążonymi estymatorami wariancji σ^2. Stosunki

$$F_R = \frac{s_R^2}{s_0^2} \tag{7.38}$$

$$F_C = \frac{s_c^2}{s_0^2} \tag{7.39}$$

powinny (przy założeniu braku efektów kolumn i wierszy) być małe, gdyż wówczas zmienność byłaby spowodowana tylko efektami losowymi (resztowymi), nie wyjaśnionymi przez zmienność między wierszami i między kolumnami. Hipotezę zerową testujemy porównując wartości stosunków F_R i F_C z wartościami krytycznymi rozkładu F Snedecora przy odpowiedniej liczbie stopni swobody i odrzucając hipotezę o braku efektów wierszy i kolumn, gdy obliczone stosunki przekraczają wartości krytyczne, czyli gdy

$$F_R \geq \alpha F^{(r-1)}_{[(r-1)(c-1)]} \qquad \text{i/lub} \qquad F_C \geq \alpha F^{(c-1)}_{[(r-1)(c-1)]}$$

Tabela 7.10 przedstawia tablicę analizy wariancji w klasyfikacji podwójnej.

Tabela 7.10

Tablica analizy wariancji w klasyfikacji podwójnej w układzie bloku zrandomizowanego

Źródło zmienności	Suma kwadratów	Liczba stopni swobody	Średni kwadrat	Stosunek wariancji
Między wierszami	$SKMW$	$r-1$	$s_R^2 = \frac{SKMW}{r-1}$	$F_R = \frac{s_R^2}{s_0^2}$
Między kolumnami	$SKMK$	$c-1$	$s_C^2 = \frac{SKMK}{c-1}$	$F_C = \frac{s_C^2}{s_0^2}$
Reszta	SKR	$(r-1)(c-1)$	$s_0^2 = \frac{SKR}{(r-1)(c-1)}$	
Całkowita	SK	$rc-1 = N-1$		

Przykład 7.3

Przeprowadzamy analizę wariancji dla danych już wcześniej analizowanych w przykładzie 7.1. Osocze do badań było pobierane od dziesięciu pacjentów. Wyniki pomiarów czasu krzepnięcia osocza dla określonego pacjenta umieszczono w jednym wierszu tabeli. Kolumny odpowiadają czterem metodom dokonywania oznaczeń. Dane wraz z pośrednimi wynikami obliczeń pokazano w tabeli 7.11, zaś wyniki końcowe w formie tablicy analizy wariancji przedstawia tabela 7.12.

Tabela 7.11

Czas krzepnięcia osocza (w minutach) mierzony czterema metodami u dziesięciu pacjentów

Metody / Pacjenci	1	2	3	4	R_i	R_i^2	$\sum_j y_{ij}^2$	$\bar{y}_i$
1	9,1	10,0	10,0	10,9	40,0	1600,00	401,62	10,00
2	8,9	10,2	9,9	11,1	40,1	1608,01	404,47	10,02
3	8,4	9,8	9,8	12,2	40,2	1616,04	411,48	10,05
4	12,8	11,6	12,9	14,4	51,7	2672,89	672,17	12,93
5	8,7	9,5	11,2	9,8	39,2	1536,64	387,42	9,80
6	9,2	9,2	9,9	12,0	40,3	1624,09	411,29	10,08
7	7,6	8,6	8,5	8,5	33,2	1102,24	276,22	8,30
8	8,6	10,3	9,8	10,9	39,6	1568,16	394,90	9,90
9	8,9	9,4	9,2	10,4	37,9	1436,41	360,37	9,48
10	7,9	8,5	8,2	10,0	34,6	1197,16	301,90	8,65
C_j	90,1	97,1	99,4	110,2			$T = 396{,}8$	
C_j^2	8118,01	9428,41	9880,36	12144,04			$s = 4021{,}84$	
$\sum_i y_{ij}^2$	830,09	950,19	1004,08	1237,48			$\sum_j C_j^2 = 39570{,}82$	
$\bar{y}_{\cdot j}$	9,01	9,71	9,94	11,02			$\sum_i R_i^2 = 15961{,}64$	

W przykładzie 7.1, gdzie nie wprowadziliśmy klasyfikacji według pacjentów, można było z ogólnej sumy kwadratów równej 85,584 wyodrębnić składnik odpowiadający za różnice między metodami wynoszący 20,826, a pozostałe 64,758 uznać za zmienną resztową (por. tabela 7.4). Obecnie po uwzględnieniu klasyfikacji podwójnej z tej poprzedniej zmienności resztowej 64,758 wyodrębniono jeszcze wartość 54,154 odpowiedzialną za (nie interesujące nas w rzeczywistości) różnice między pacjentami, więc miarę zmienności między metodami należy odnieść do miary zmienności resztowej wynoszącej już tylko 10,604. Uzyskano znacznie silniejsze potwierdzenie różnic między metodami pomiaru czasu krzepnięcia osocza. Podejście takie — poza tym — wydaje się rozsądniejsze, bo pozwala uniezależnić się od zmienności osobniczej pacjentów, czyli mówiąc innymi słowami, uniezależnić się od niejednorodności „materiału" eksperymentu.

Tabela 7.12

Tablica analizy wariancji dla danych z tabeli 7.11

Źródło zmienności	Suma kwadratów	Liczba stopni swobody	Średni kwadrat	Stosunek wariancji	Istotność
Między pacjentami	54,154	9	6,0171	15,32	$P < 0{,}001$
Między metodami	20,826	3	6,924	17,68	$P < 0{,}001$
Reszta	10,604	27	0,3927		
Całkowita	85,584	39			

7.2.2 Test interakcji

Rozważania w punkcie poprzednim przeprowadzono zakładając możliwość zastosowania do analizowanych danych modelu liniowego wyrażonego zależnościami (7.28) i (7.29). Przykładem danych zgodnych z modelem liniowym będzie poniższa hipotetyczna tablica:

	*A*1	*A*2	*A*3
*B*1	5	10	20
*B*2	10	15	25
*B*3	25	30	40

asność, że różnice między kolumnami są stałe niezależnie od wiersza, podobnie różnice między wierszami są stałe, niezależne od kolumn. Obydwa czynniki *A* i *B* dają efekty sumujące się. Kategoria *A*3 ma większe wartości niż *A*2, zaś *A*2 większe niż *A*1. Podobne zależności zachodzą w klasyfikacji B. Jeżeli jednak zmieni się wartość np. środkowego pola tabeli i wpisze się tam 35 zamiast 15,

	A1	A2	A3
B1	5	10	20
B2	10	35	25
B3	25	30	40

to będzie to dowód na szczególną współzależność (interakcję) między kategorią $A2$ i $B2$, przy zachowaniu ogólnej tendencji większych wartości w trzeciej kolumnie i trzecim wierszu w porównaniu z drugą kolumną i drugim wierszem, itd.

W przypadkach, gdy model addytywny (7.28) nie jest odpowiedni, stosujemy model uwzględniający interakcję między efektami wierszy i kolumn, dany zależnością:

$$E(y_{ij}) = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} \tag{7.40}$$

gdzie $(\alpha\beta)_{ij}$ jest stałą odpowiedzialną za efekt interakcji, spełniającą związek

$$\sum_{ij} (\alpha\beta)_{ij} = 0$$

Analizę wariancji uwzględniającą interakcję można przeprowadzić wówczas, gdy w każdej podgrupie klasyfikacyjnej (w każdym polu tabeli leżącym w i-tym wierszu i j-tej kolumnie) mamy nie jedną, lecz więcej obserwacji. W przypadku jednej obserwacji wewnątrz pola tabeli nie można odróżnić efektu interakcji od zmienności resztowej.

Będziemy rozważać przypadek, gdy w każdej podgrupie klasyfikacyjnej jest jednakowa liczba obserwacji. Oznaczmy tę liczbę replikacji obserwacji w każdym polu tablicy o wymiarach $r \times c$ przez n. Sumy marginalne R_i oraz C_i będziemy podobnie jak poprzednio wykorzystywać dla utworzenia sum kwadratów odpowiedzialnych za efekty główne (efekt wierszy i efekt kolumn). Sumy T_{ij} (sumy obserwacji wewnątrz pól tabeli) posłużą m.in. do oszacowania efektu interakcji, zaś znajomość wartości poszczególnych obserwacji wykorzystujemy m.in. do oszacowania zmienności resztowej — nie wyjaśnionej efektami głównymi i efektem interakcji. Będziemy przyjmować oznaczenia z tabeli 7.13, przy czym y_{ijp} będzie wartością p-tej obserwacji w i-tym wierszu i j-tej kolumnie. Założenia w teście analizy wariancji uwzględniającym interakcje są podobne jak poprzednio. Zakłada się w szczególności normalność rozkładów w podgrupach i stałą wariancję. Hipoteza zerowa, poza równością średnich w wierszach i równością średnich w kolumnach, zakłada teraz także brak interakcji (addytywność). Przeprowadzając test analizy wariancji rozbijamy ogólną sumę kwadratów (SK) odchyleń od średniej $\bar{y}$ na cztery składniki odpowiadające:

Tabela 7.13

Analiza wariancji w klasyfikacji podwójnej z uwzględnieniem efektu interakcji — oznaczenia

kolumny / wiersze	1	2	... j ...	c	Suma
1	T_{11}	T_{12}	T_{1j}	T_{1c}	R_1
2	T_{21}	T_{22}	T_{2j}	T_{2c}	R_2
:			$y_{ij1} \cdot y_{ij2}$		
i	T_{i1}	T_{i2}	$\ldots \cdot y_{ijn}$	T_{ic}	R_i
:			T_{ij}		
r	T_{r1}	T_{r2}	T_{rj}	T_{rc}	R_r
suma	C_1	C_2	C_j	C_c	T

$$T_{ij} = \sum_{p=1}^{n} y_{ijp} \qquad C_j = \sum_{i=1}^{r} T_{ij} \qquad R_i = \sum_{j=1}^{c} T_{ij}$$

$$T = \sum_i R_i = \sum_i C_j = \sum_{i,j} T_{ij} = \sum_{i,j,p} y_{ijp} \qquad s = \sum_{i,j} y_{ij}^2$$

$$N = r \cdot c \cdot n \qquad \bar{y} = \frac{T}{N}$$

— efektowi wierszy (*SKMW*)
— efektowi kolumn (*SKMK*)
— efektowi interakcji (*SKI*)
— zmienności resztowej (*SKR*)

$$SK = SKMW + SKMK + SKI + SKR$$

Poszczególne sumy kwadratów liczymy ze wzorów:

$$SK = S - \frac{T^2}{N} \tag{7.41}$$

$$SKMW = -\frac{\sum_i R_i^2}{nc} - \frac{T^2}{N} \tag{7.42}$$

$$SKMK = -\frac{\sum_j C_j^2}{nr} - \frac{T^2}{N} \tag{7.43}$$

$$SKMW = -\frac{\sum_{i,j} T_{ij}^2}{n} - \frac{T^2}{N} - SKMW - SKMK \tag{7.44}$$

$$SKR = SK - SKMW - SKMK - SKI \tag{7.45}$$

Dalsze postępowanie polegające na utworzeniu odpowiednich średnich kwadratowych i stosunków wariancji F ilustruje tabela 7.14.

Tabela 7.14

Tablica analizy wariancji w klasyfikacji podwójnej z uwzględnieniem interakcji

Źródło zmienności	Suma kwadratów	Liczba stopni swobody	Średni kwadrat	Stosunek wariancji
Między wierszami	$SKMW$	$r - 1$	$s_R^2 = \frac{SKMW}{r-1}$	$F_R = \frac{s_R^2}{s_0^2}$
Między kolumnami	$SKMK$	$c - 1$	$s_C^2 = \frac{SKMK}{c-1}$	$F_C = \frac{s_C^2}{s_0^2}$
Interakcja	SKI	$(r - 1)(c - 1)$	$s_I^2 = \frac{SKI}{(r-1)(c-1)}$	$F_I = \frac{s_I^2}{s_0^2}$
Reszta	SKR	$N - rc$	$s_0^2 = \frac{SKR}{N-rc}$	
Całkowita	SK	$N - 1$		

Stosunki wariancji testujemy wykorzystując w znany sposób rozkład F Snedecora. Niektórzy autorzy (np. [Blalock]) polecają rozpocząć testowanie od ilorazu wariancji F_I. Przy braku podstaw do odrzucenia hipotezy o addytywności zalecają oni sumę kwadratów interakcji dodać do składnika resztowego zmieniając odpowiednio liczbę stopni swobody.

$$SKR' = SKR + SKI \tag{7.46}$$

$$(s_0')^2 = \frac{SKR'}{N - r - c + 1} \tag{7.47}$$

i użyć tak zmodyfikowanego (wzór (7.47)) oszacowania resztowego jako mianownika stosunków F dla badania efektów głównych.

Gdyby zaś interakcja okazała się istotna, można obliczyć dla każdego pola tabeli wartość resztową

$$d_{ij} = y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y} \tag{7.48}$$

(gdzie: $\bar{y}_{ij} = \frac{T_{ij}}{n}$, $\bar{y}_{.j} = \frac{C_j}{rn}$, $\bar{y}_{i.} = \frac{R_i}{cn}$, $\bar{y} = \frac{T}{N}$)

która to wartość d_{ij} wskazuje odchylenie średniej obserwacji w podgrupie od wartości spodziewanej, przy przyjęciu modelu addytywnego, tzn. wartości, którą powinna przyjąć średnia przy braku interakcji. W ten sposób można zlokalizować pola tabeli które wnoszą największy wkład w ową interakcję. Konieczność teoretycznego wyjaśnienia interakcji może doprowadzić do spostrzeżeń, które będą najważniejszymi wynikami całych badań.

Przykład 7.4

Rozważmy część danych z tabeli 7.11 (zaznaczoną linią przerywaną). Załóżymy, że dane z tabeli 7.11 były średnimi z trzech obserwacji. Uwzględniając wszystkie trzy obserwacje dla każdej z trzech metod u trzech pacjentów otrzymamy tabelę 7.15. Dla danych tam zawartych przeprowadzimy test analizy wariancji. Mamy:

$r = 3$ $c = 3$ $n = 3$ $N = 27$ $S = 2526{,}37$

Dalsze wyniki przedstawia tabela 7.16. Efekty dotyczące pacjentów i metod okazały się wysoce istotne, zaś interakcja nie jest istotna na poziomie 0,05, choć stosunek F_I nie jest mały.

Tabela 7.15

Czas krzepnięcia osocza (w minutach) trzech pacjentów przy zastosowaniu trzech metod pomiaru i przy trzech replikacjach każdej kombinacji pacjent-metoda

Metody / Pacjenci	2	3	4	Sumy
8	T_{11} = 30,9 10,2 = y_{111} 10,5 = y_{112} 10,2 = y_{113} S_{11} = 318,33	29,4 9,9 9,5 10,0 288,26	32,7 11,3 10,7 10,7 356,67	R_1 = 93,0
9	28,2 9,6 9,0 9,6 265,32	27,6 9,1 9,1 9,4 253,98	31,2 10,3 10,7 10,2 324,62	R_2 = 87,0
10	25,5 9,0 8,1 8,4 217,17	24,6 8,6 8,0 8,0 201,96	30,0 9,8 10,1 10,1 300,06	R_3 = 80,1
Sumy	C_1 = 84,6	C_2 = 81,6	C_3 = 93,9	T = 260,1
$s_{ij} = \sum_{p=1}^{n} y_{ijp}^2$				

$$2{,}08 = F_I < {}_{0{,}05}F^{(4)}_{(18)} = 2{,}93$$

Poleca się czytelnikowi dokończenie obliczeń według schematu proponowanego w tekście powyżej.

Tabela 7.16

Tablica analizy wariancji dla danych z tabeli 7.15

Źródło zmienności	Suma kwadratów	Liczba stopni swobody	Średni kwadrat	Stosunek wariancji	Istotność
Między pacjentami	9,26	2	4,63	52,08	$P < 0{,}001$
Między metodami	9,14	2	4,57	51,41	$P < 0{,}001$
Interakcja	0,74	4	0,185	2,08	nieistotne
Reszta	1,60	18	0,0889		
Całkowita	20,74	26			

7.3 Analiza wariancji w klasyfikacji hierarchicznej

W podrozdziale 7.2 poznaliśmy sposób analizy obserwacji ilościowych poddanych dwóm prostym klasyfikacjom jakościowym. Często zachodzi potrzeba rozpatrzenia wyników obserwacji w klasyfikacji hierarchicznej. Rozważa się więc pierwotną klasyfikację na I grup. Każdą z tych grup „najwyższego" poziomu można podzielić na J grup następnego poziomu, te grupy z kolei na K grup jeszcze niższego poziomu, itd. W każdej z pojedynczych grup najniższego poziomu dysponuje się pewną liczbą obserwacji. Tutaj analiza wariancji ma udzielić odpowiedzi na pytanie: czy obserwowane zróżnicowanie średnich między grupami pewnego poziomu hierarchi jest uwarunkowane tylko zmiennością losową wewnątrz poszczególnych grup tego poziomu, czy też ma charakter istotnego zróżnicowania nie dającego się wyjaśnić efektami losowymi niższego szczebla. W związku z tym w analizie wariancji porównujemy oszacowanie wariancji między grupami pewnego poziomu A w stosunku do oszacowania wariancji między grupami niższego poziomu B, ale tylko w ramach grup poziomu A.

Przedstawiamy teraz sposób analizy danych ilościowych z zastosowaniem analizy wariancji w przypadku dwupoziomowej klasyfikacji hierarchicznej. Będziemy rozważać podział obserwacji na I grup A_i wyższego poziomu A. Każda z grup A_i ($i = 1, ..., I$) dzieli się dalej na J grup B_{ij} niższego poziomu B, każda zaś z grup B_{ij} ($j = 1, ..., J$)

zawiera K obserwacji y_{ijk} (k = 1, ..., K). Klasyfikację taką pokazano w tabeli 7.17. Przyjmując oznaczenia:

$$T_{ij} = \sum_{k=1}^{K} y_{ijk} \qquad S_{ij} = \sum_{k=1}^{K} y_{ijk}^2$$

$$T_i = \sum_{j=1}^{J} T_{ij} \qquad S_i = \sum_{j=1}^{J} S_{ij}$$

$$T = \sum_{i=1}^{I} T_{ij} \qquad S = \sum_{i=1}^{I} S_i$$

$$N = I * J * K$$

możemy całkowitą sumę kwadratów SK przedstawić jako sumę kwadratów SKA dla poziomu A oraz sumę kwadratów SKB dla poziomu B

$$SK = SKA + SKB$$

gdzie

$$SKA = \frac{\sum_i T_i^2}{JK} - \frac{T^2}{N} \tag{7.49}$$

$$SKB = \frac{\sum_{i,j} T_{i,j}^2}{K} - \frac{T^2}{N} \tag{7.50}$$

$$SK = S - \frac{T^2}{N} \tag{7.51}$$

Taki podział jednak nie umożliwia bezpośrednich porównań między sąsiednimi poziomami hierarchi. Korzystne jest więc zastąpić go podziałem:

$$SK = SKA + SKBA + SKOB \tag{7.53}$$

przy czym $SKBA$ jest sumą kwadratów między podgrupami B_{ij}, ale tylko w obrębie odpowiednich grup A_i i wyraża się zależnością:

Tabela 7.17

Dwupoziomowa klasyfikacja hierarchiczna — oznaczenia

Poziom A	A_1	...	A_i	...	A_I
Poziom B	B_{11} ... B_{1j} ... B_{1J}		B_{i1} .. . B_{ij} . .. B_{iJ}		B_{I1} ... B_{Ij} ... B_{IJ}
Obserwacje wewnątrz poziomu B			y_{ij1} ⋮ y_{ijk} ⋮ y_{ijK}		
Sumy wewnątrz podgrup B_{ij}			T_{ij} S_{ij}		
Sumy wewnątrz grup A_i	T_1 S_1	...	T_i S_i	...	T_I S_I
Sumy całkowite	T S				

$$SKBA = SKB - SKA$$

zaś $SKOB$ jest sumą kwadratów między obserwacjami w obrębie podgrup B_{ij}:

$$SKOB = SK - SKB \tag{7.54}$$

Tablicę analizy wariancji dla klasyfikacji hierarchicznej pokazano w tabeli 7.18. Warto zwrócić uwagę, że tutaj każdorazowo tworzymy stosunki wariancji porównując średni kwadrat z sąsiednim średnim kwadratem znajdującym się tuż pod nim. Odpowiada to porównywaniu zmienności między sąsiednimi poziomami hierarchi przy kolejnym schodzeniu coraz niżej. Stosunki wariancji testujemy w znany sposób, pamiętając że F_1 charakteryzuje się $I - 1$ i $I \cdot (J - 1)$ stopniami swobody, zaś F_2 odpowiednio $I \cdot (J - 1)$ i $IJ \cdot (K - 1)$ stopniami swobody.

Tabela 7.18

Tablica analizy wariancji w klasyfikacji hierarchicznej

Źródło zmienności	Suma kwadratów	Liczba stopni swobody	Średni kwadrat	Stosunek wariancji
Między grupami A_i poziomu A	SKA	$I - 1$	$s_A^2 = \frac{SKA}{I-1}$	$F_1 = \frac{s_A^2}{s_{BA}^2}$
Między podgrupami B_{ij} w obrębie grup A_i	$SKBA$	$I(J - 1)$	$s_{BA}^2 = \frac{SKBA}{I(J-1)}$	$F_2 = \frac{s_{BA}^2}{s_{OB}^2}$
Między obserwacjami w obrębie podgrup B_{ij}	$SKOB$	$IJ(K - 1)$	$s_{OB}^2 = \frac{SKOB}{IJ(K-1)}$	
Całkowita	SK	$N - 1$		

Przykład 7.5 (w/g [Parker])

Badano wpływ światła o różnym natężeniu na liczbę chloroplastów w komórkach liści mchu *Funaria hygrometrica*. Ostatecznym celem badania miało być ustalenie średniej liczby chloroplastów w komórce przy określonym natężeniu światła. Planując eksperyment należało odpowiedzieć na pytanie: ile komórek należy uwzględnić na jednym liściu, ile liści badać na jednej roślinie i ile roślin brać pod uwagę. Zebrane dane mające ułatwić odpowiedź na to pytanie przedstawia tabela 7.19. Wstępnie przebadano 4 rośliny, 5 liści na każdej roślinie i 10 losowo wybranych komórek w każdym liściu.

Tabela 7.19

Wyniki badań liczby chloroplastów w komórkach mchu

Roślina (A_i)	Roślina A_1					Roślina A_2				
Liść (B_{ij})	1	2	3	4	5	1	2	3	4	5
T_{ij}	134	126	127	135	132	145	130	148	139	140
S_{ij}	1814	1622	1625	1867	1792	2141	1712	2189	2039	1992
T_i	654					702				
Roślina (A_i)	Roślina A_3					Roślina A_4				
Liść (B_{ij})	1	2	3	4	5	1	2	3	4	5
T_{ij}	146	147	143	152	149	140	121	134	146	145
S_{ij}	2174	2167	2079	2314	2239	1980	1495	1818	2132	2188
T_i	737					686				

Po przeprowadzeniu obliczeń mamy:

$T = 2779$

$S = 39388$

$\frac{T^2}{N} = 38614{,}21$

$SK = 773{,}79$

$SKB = 143{,}49$

$SKBA = 72{,}00$

$SKA = 71{,}49$

$SKOB = 630{,}00$

Tablicę analizy wariancji dla rozważanych obserwacji przedstawia tabela 7.20. Widać, że średni kwadrat dla porównań między liśćmi nie jest istotnie większy od średniego kwadratu dla porównań między pomiarami w obrębie liści (F_2), zaś średni kwadrat dla porównań między roślinami w stosunku do średniego kwadratu dla porównań między liścmi w obrębie roślin wskazuje na istotny wzrost (F_1).

Tabela 7.20

Tablica analizy wariancji dla wyników badań liczby chloroplastów w komórkach mchu (por. tab. 7.19)

Źródło zmienności	Suma kwadratów	Liczba stopni swobody	Średni kwadrat	Stosunek wariancji	Istotność
Między roślinami	71,49	3	23,83	5,50	$P < 0{,}025$
Między liśćmi w obrębie roślin	72,00	16	4,50	1,29	nieistotne
Między komórkami w obrębie liści	630,30	180	3,50		
Całkowita	733,79	199			

Na podstawie tych wyników można sugerować, że najlepszą skuteczność w estymacji średniej liczby chloroplastów w świetle o określonym natężeniu otrzymalibyśmy stosując przy każdym natężeniu światła dużą liczbę próbek roślinnych i licząc chloroplasty w jednej komórce lub w jednym liściu każdej z roślin.

7.4 Analiza wariancji w schemacie kwadratu łacińskiego

Załóżmy, że chcemy badać wyniki leczenia a metodami w układzie w którym mamy do czynienia z dwoma innymi źródłami zmienności, każde także o a kategoriach. Decydując się na dokonanie po jednej obserwacji dla każdej metody i każdego czynnika zmienności musielibyśmy wykonać a^3 obserwacji (byłby to tzw. schemat trójczynnikowy stanowiący uogólnienie problemu omówionego w podrozdziale 7.2). Możemy zdecydować się na inny sposób postępowania i zaplanować tak doświadczenie, aby każdą metodę leczenia stosować tylko jeden raz dla każdej kategorii klasyfikacji pierwszej i także tylko raz dla każdej kategorii klasyfikacji drugiej. Schemat takiego postępowania dla $a = 6$ przedstawiono w tabeli 7.21.

Tabela 7.21

Przykład kwadratu łacińskiego o boku równym 6

		Kolumny					
		1	2	3	4	5	6
Wiersze	1	*C*	*E*	*D*	*A*	*F*	*B*
	2	*D*	*B*	*F*	*E*	*C*	*A*
	3	*A*	*C*	*E*	*F*	*B*	*D*
	4	*F*	*A*	*C*	*B*	*D*	*E*
	5	*B*	*D*	*A*	*C*	*E*	*F*
	6	*E*	*F*	*B*	*D*	*A*	*C*

Metody leczenia oznaczono tutaj dużymi literami łacińskimi od *A* do *F*, a wtórne klasyfikacje odpowiadają wierszom i kolumnom. W każdym wierszu i każdej kolumnie dana litera występuje dokładnie jeden raz. Przy takim schemacie postępowania zwanym kwadratem łacińskim dokonuje się a^2 obserwacji, ponieważ w każdym polu tabeli jest tylko jeden z a sposobów leczenia.

Kwadraty łacińskie tworzy się w ten sposób, że zapełnia się cyklicznie wiersze tabeli, np.

A	*B*	*C*	*D*	*E*	*F*
F	*A*	*B*	*C*	*D*	*E*
E	*F*	*A*	*B*	*C*	*D*
D	*E*	*F*	*A*	*B*	*C*
C	*D*	*E*	*F*	*A*	*B*
B	*C*	*D*	*E*	*F*	*A*

a następnie w sposób losowy przestawia się wiersze i kolumny.

Przy analizie kwadratu łacińskiego przyjmuje się model addytywny

$$y_{ijk} = \mu + \alpha_i + \beta_i + \gamma_k + \varepsilon_{ijk} \tag{7.55}$$

gdzie

y_{ijk} — obserwacja w i-tym wierszu, j-tej kolumnie i dla k-tej metody leczenia

μ — średnia ogólna

α_i, β_j, γ_k — stałe odpowiadające za efekty główne wierszy, kolumn i metod

ε_{ijk} — składnik losowy o rozkładzie normalnym, zerowej wartości średniej i wariancji σ^2.

Model taki nie uwzględnia interakcji, gdyż w jednym polu tabeli mamy dokładnie jedną obserwację. Przyjmujemy oznaczenia jak w tabeli 7.22. Całkowitą sumę kwadratów odchyleń od średniej

Tabela 7.22

Analiza wariancji w schemacie kwadratu łacińskiego — oznaczenia

		Kolumny								Metody	
		1	2	...	j	...	a	Sumy	Średnie	Sumy	Średnie
Wiersze	1							R_1	$\bar{y}_{1\cdot\cdot}$	T_1	$\bar{y}_{\cdot\cdot 1}$
	2							R_2	$\bar{y}_{2\cdot\cdot}$	T_2	$\bar{y}_{\cdot\cdot 2}$
	:							:	:	:	:
	i				y_{ijk}			R_i	$\bar{y}_{i\cdot\cdot}$	T_k	$\bar{y}_{\cdot\cdot k}$
	:							:	:	:	:
	a							R_a	$\bar{y}_{a\cdot\cdot}$	T_a	$\bar{y}_{\cdot\cdot a}$
	Sumy	C_1	C_2	...	C_j	...	C_a	T		T	
	Średnie	$\bar{y}_{\cdot 1\cdot}$	$\bar{y}_{\cdot 2\cdot}$	...	$\bar{y}_{\cdot j}$	...	$\bar{y}_{\cdot a}$		$\bar{y}$		$\bar{y}$
$T = \sum_i R_i = \sum_j C_j = \sum_k T_k$					$N = a^2$				$\bar{y} = \frac{T}{N}$		

$$SK = \sum \left(\bar{y}_{ijk} - \bar{y}\right)^2 = \sum_{i,j} y_{ijk}^2 - \frac{T^2}{a^2} \tag{7.56}$$

można rozdzielić na składniki odpowiedzialne za trzy efekty główne: wierszy (*SKMW*), kolumn (*SKMK*) i metod (*SKMM*) oraz wyodrębnić składnik resztowy (*SKR*):

$$SKMW = \sum_i \left(\bar{y}_{i\cdot\cdot} - \bar{\bar{y}}\right)^2 = \sum_i \frac{R_i^2}{a} - \frac{T^2}{a^2} \tag{7.57}$$

$$SKMK = \sum_j \left(\bar{y}_{\cdot j\cdot} - \bar{\bar{y}}\right)^2 = \sum_j \frac{C_j^2}{a} - \frac{T^2}{a^2} \tag{7.58}$$

$$SKMM = \sum_k \left(\bar{y}_{\cdot\cdot k} - \bar{\bar{y}}\right)^2 = \sum_k \frac{T_k^2}{a} - \frac{T^2}{a^2} \tag{7.59}$$

$$SKR = \sum \left(y_{ijk} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} - \bar{y}_{\cdot\cdot k} + 2\bar{\bar{y}}\right) = SK - SKMW - SKMK - SKMM \tag{7.60}$$

Tabela 7.23

Tablica analizy wariancji w schemacie kwadratu łacińskiego

Źródło zmienności	Suma kwadratów	Liczba stopni swobody	Średni kwadrat	Stosunek wariancji
Wiersze	$SKMW$	$a - 1$	$s_R^2 = \frac{SKMW}{a-1}$	$F_R = \frac{s_R^2}{s_0^2}$
Kolumny	$SKMK$	$a - 1$	$s_C^2 = \frac{SKMK}{a-1}$	$F_C = \frac{s_C^2}{s_0^2}$
Metody	$SKMM$	$a - 1$	$s_T^2 = \frac{SKMM}{a-1}$	$F_T = \frac{s_T^2}{s_0^2}$
Reszta	SKR	$(a-1)(a-2)$	$s_0^2 = \frac{SKR}{(a-1)(a-2)}$	
Ogółem	SK	$a^2 - 1$		

Porównując oszacowania wariancji efektów głównych z wariancją resztową otrzymujemy trzy stosunki wariancji, które testujemy w zwykły sposób przy wykorzystaniu rozkładu F Snedecora (por. tabela 7.23).

Przykład 7.6

Sześciu różnym królikom wstrzykiwano w sześć różnych miejsc na skórze grzbietu pewien preparat. Stosowano sześć różnych sekwencji kolejnych szczepień tego samego zwierzęcia.

Tabela 7.24

Wyniki pomiarów wielkości pęcherza (w cm^3) po wstrzyknięciu preparatu w skórę grzbietu sześciu królików w miejscach a — f, kolejność wstrzyknięć A — F

		Zwierzęta							
		1	2	3	4	5	6	Ogółem	Średnia
Miejsce	*a*	*C* 7,9	*E* 8,7	*D* 7,4	*A* 7,4	*F* 7,1	*B* 8,2	46,7	7,783
	b	*D* 6,1	*B* 8,2	*F* 7,7	*E* 7,1	*C* 8,1	*A* 5,9	43,1	7,183
	c	*A* 7,5	*C* 8,1	*E* 6,0	*F* 6,4	*B* 6,2	*D* 7,5	41,7	6,950
	d	*F* 6,9	*A* 8,5	*C* 6,8	*B* 7,7	*D* 8,5	*E* 8,5	46,9	7,817
	e	*B* 6,7	*D* 9,9	*A* 7,3	*C* 6,4	*E* 6,4	*F* 7,3	44,0	7,333
	f	*E* 7,3	*F* 8,3	*B* 7,3	*D* 5,8	*A* 6,4	*C* 7,7	42,8	7,133
	Ogółem	42,4	51,7	42,5	40,8	42,7	45,1	265,2	
	Średnia	7,067	8,617	7,083	6,800	7,117	7,517		7,367
Kolejność		A	B	C	D	E	F	$\sum y_{ijk}^2 = 1984$	
	Ogółem	43,0	44,3	45,0	45,2	44,0	43,7		
	Średnia	7,167	7,383	7,500	7,533	7,333	7,283		

Po upływie pewnego czasu mierzono powierzchnię pęcherzy powstałych w miejscach szczepień. Doświadczenie przeprowadzono w schemacie kwadratu łacińskiego. Jego wyniki przedstawia tabela 7.24. Analiza wariancji uzyskanych danych (por. tabela 7.25) wykazała, że jedynym istotnym efektem głównym są różnice zachodzące w obrębie zwierząt. W związku z tym efekty miejsca i kolejności można pominąć i dalsze rozważania przeprowadzać wykorzystując metody stosowane dla szczegółowej analizy różnic między średnimi w klasyfikacji pojedynczej.

Tabela 7.25

Tablica wariancji dla danych w schemacie kwadratu łacińskiego (z tabeli 7.24)

Źródło zmienności	Suma kwadratów	Liczba stopni swobody	Średni kwadrat	Stosunek wariancji	Istotność
Miejsca	3,8332	5	0,7667	1,17	nieistotne
Zwierzęta	12,8333	5	2,5667	3,91	$P < 0,05$
Kolejność	0,5632	5	0,1106	0,17	nieistotne
Reszta	13,1302	20	0,6565		
Ogółem	30,3599	35			

Gdy model addytywny (7.55) w schemacie kwadratu łacińskiego nie jest odpowiedni, można w pewnych przypadkach szacować istotność efektu interakcji. Do tego jednak niezbędne jest posiadanie w każdym polu tabeli r replikacji obserwacji ($r \geq 2$). Dalsze informacje na ten temat można znaleźć w literaturze [Armitage].

Istnieje wiele różnych dalszych niekompletnych schematów planowania doświadczeń. Gdy chcemy badać jakieś zjawisko o charakterze ilościowym w klasyfikacji poczwórnej (każda klasyfikacja o a kategoriach) i nie możemy z jakiś względów stosować pełnego schematu czynnikowego, wymagającego a^4 danych, możemy skorzystać np. z kwadratów grecko-łacińskich (por. tab. 7.26). Tutaj dwie klasyfikacje odpowiadają wierszom i kolumnom, zaś następne dwie — literom greckim i łacińskim. Można zauważyć, że zarówno litery łacińskie, jak i litery greckie tworzą kwadraty łacińskie i że każda litera łacińska towarzyszy każdej z liter greckich dokładnie jeden raz.

Tabela 7.26

Przykład kwadratu grecko-łacińskiego o boku równym 5

		Kolumny				
		1	2	3	4	5
Wiersze	1	$A\alpha$	$C\gamma$	$E\varepsilon$	$B\beta$	$D\delta$
	2	$B\delta$	$D\alpha$	$A\gamma$	$C\varepsilon$	$E\beta$
	3	$C\beta$	$E\delta$	$B\alpha$	$D\gamma$	$A\varepsilon$
	4	$D\varepsilon$	$A\beta$	$C\delta$	$E\alpha$	$B\gamma$
	5	$E\gamma$	$D\varepsilon$	$D\beta$	$A\delta$	$C\alpha$

8. REGRESJA I KORELACJA

8.1 Regresja liniowa. Współczynnik korelacji

Przedstawimy teraz sposób analizy innego typu danych ilościowych. Będziemy dla każdej badanej „jednostki" dysponowali obserwacjami dwóch zmiennych ilościowych x i y. Celem będzie zbadanie związku między tymi dwiema zmiennymi. Na przykład badając przyczyny nieregularnego wzrostu drzew na pewnej plantacji świerka *Picea sitchensis* można wybrać w sposób losowy pewną liczbę drzew, określić dokładnie ich wysokość (po uprzednim ścięciu), a następnie za pomocą analizy chemicznej wyznaczyć zawartość substancji odżywczych w igłach świeżych pędów. W takim badaniu jedną zmienną (np. zmienną y) będzie wysokość drzewa w metrach, a drugą zmienną (czyli x) zawartość azotu w igłach wyrażona w % azotu na jednostkę suchej masy. Przedstawiane poniżej metody pozwolą uzyskać pewne informacje o związku między dwiema zmiennymi, gdy rozkład jednej zmiennej (np. zmiennej y) związany jest z wartościami drugiej zmiennej (czyli x). Nie znaczy to, że jedna zmienna jest przyczyną drugiej, nie mówimy więc o związku przyczynowo-skutkowym. Na ogół jednak tak określamy zmienną x i zmienną y, żeby z ewentualną zależnością $y = f(x)$ móc w pewnych okolicznościach łączyć związek przyczynowo-skutkowy. W naszym przykładzie można by się dopatrywać zależności wysokości drzewa od poziomu zawartości substancji odżywczych w młodych pędach.

Badanie związku dwóch zmiennych ilościowych prowadzimy na ogół w celu:

— uzyskania liczbowych miar pewnych podstawowych cech związku,

— dostarczenia możliwości prognozowania (predykcji) wartości jednej ze zmiennych, gdy druga jest znana,

— stwierdzenia, czy obserwowany kierunek trendu jest istotny.

Wprowadzimy teraz pojęcie funkcji regresji. Załóżmy, że dysponujemy obserwacjami zmiennych x i y dla dużej liczby jednostek. Interesuje nas, jakiej przeciętnej zmianie ulega y, gdy x przyjmuje różne wartości. Jeżeli będziemy rozpatrywać pewną konkretną wartość x, to odpowiadające temu x wartości y będą zmienną losową. Wartość oczekiwaną takiej zmiennej losowej warunkowej oznaczymy $E(y|x)$. Wartość oczekiwana $E(y|x)$ zależy od x. Zależność tę nazywamy funkcją regresji zmiennej y względem zmiennej x. Wykres zależności $E(y|x)$ od x nazywany jest *krzywą regresji*.

Szczególnie często rozważamy krzywe regresji będące liniami prostymi

$$E(y|x) = A + B \cdot x \tag{8.1}$$

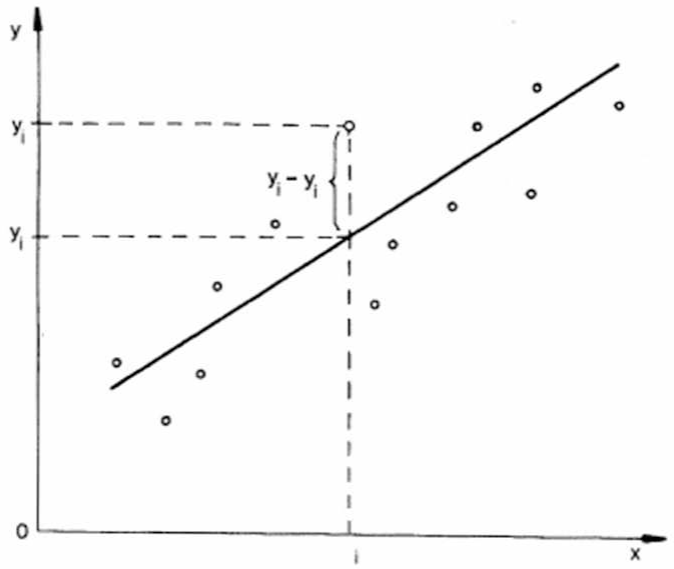
Nazywamy je wówczas prostymi regresji. Będą nas dalej interesować proste regresji charakteryzujące się stałym rozrzutem zmiennej $y|x$, niezależnym od wartości x. Takie proste regresji będziemy nazywać homoskedastycznymi. Kolejnym przyjmowanym przez nas założeniem będzie normalność rozkładu zmiennej y dla każdego x. Rozkład y ma być więc rozkładem normalnym ze średnią $E(y|x) = A + B \cdot x$ i stałą wariancją σ^2.

Mamy n par obserwacji (x_i, y_i). Parametry A i B prostej regresji szacuje się w taki sposób, aby otrzymane estymatory a i b minimalizowały sumę kwadratów $\sum (y_i - Y_i)^2$, gdzie Y_i są wartościami wyznaczonymi przez oszacowane równanie regresji:

$$Y_i = a + b \cdot x_i \tag{8.2}$$

Różnice $y_i - Y_i$ są odległościami obserwowanych wartości y_i od wartości teoretycznych Y_i, czyli od linii regresji (por. rys. 8.1). Estymatory a i b wyrażają się wzorami:

$$a = \overline{y} + b \cdot \overline{x} \tag{8.3}$$



Rys. 8.1 Prosta regresji wyznaczona metodą najmniejszych kwadratów.

$$b = \frac{\sum_i \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sum_i \left(x_i - \bar{x}\right)^2} \tag{8.4}$$

Do praktycznych obliczeń dogodniejszy jest jednak inny wzór na b — równoważny (8.4):

$$b = \frac{\sum_i x_i y_i - n\,\bar{x}\,\bar{y}}{\sum_i x_i^2 - n\,\bar{x}} \tag{8.5}$$

Jeżeli można także x traktować jako zmienną losową i spełnione są założenia o normalności rozkładu dwuwymiarowego (x, y), to nic nie stoi na przeszkodzie, aby rozważać prostą regresji x względem y

$$x_i = a_{xy} + b_{xy} \cdot y_i \tag{8.6}$$

gdzie:

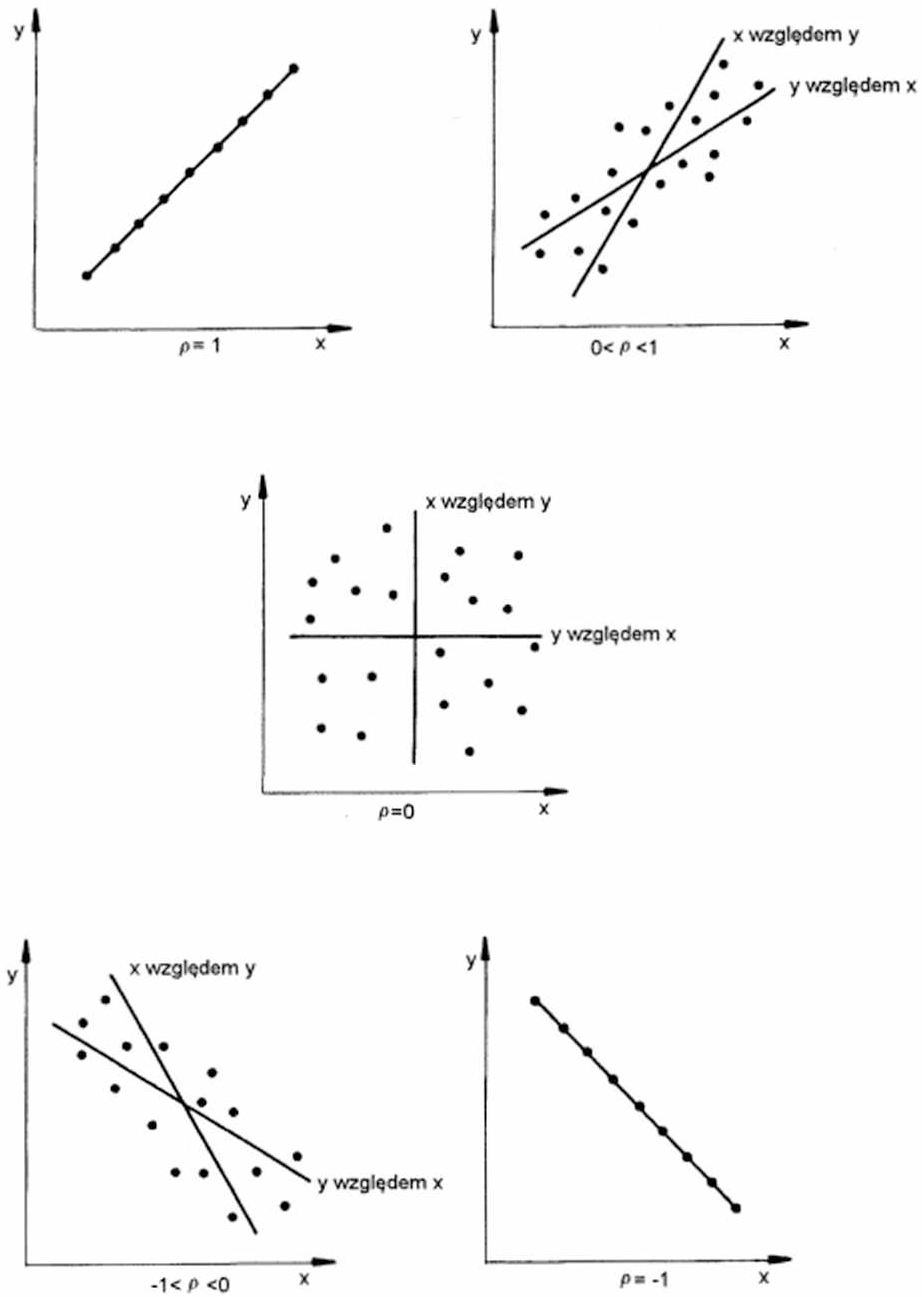
$$b_{xy} = \frac{\sum_i \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sum_i \left(y_i - \bar{y}\right)^2} \tag{8.7}$$

W ogólnym przypadku obie proste regresji (8.2) i (8.6) różnią się od siebie. Ich punktem przecięcia jest punkt $(\bar{x}, \bar{y})$.

Przy badaniu związku między dwiema cechami mierzalnymi posługujemy się dwoma pojęciami: regresji i korelacji. Pojęcie regresji już znamy — związane jest ono z kształtem zależności. Pojęcie korelacji zaś używane jest do badania siły tej zależności. Jeżeli rozważana zależność jest liniowa, to najlepszą oceną korelacji jest współczynnik korelacji ρ definiowany jako

$$\rho = \frac{\operatorname{cov}(x, y)}{\sigma_x \sigma_y} \tag{8.8}$$

gdzie $\operatorname{cov}(x, y)$ jest kowariancją zmiennych x i y, zaś σ_x i σ_y są odpowiednimi odchyleniami standardowymi. Współczynnik ρ przyjmuje wartości z przedziału <–1; 1>, przy czym gdy $\rho = -1$ lub $\rho = 1$ między zmiennymi istnieje ścisła zależność liniowa, zaś gdy $\rho = 0$ zmienne są nieskorelowane (nie można mówić o zależności liniowej). Znak współczynnika korelacji jest identyczny ze znakiem współczynnika kierunkowego B prostej regresji (por. rys. 8.2).



Rys. 8.2 Proste regresji dla różnych wartości współczynnika korelacji.

Estymatorem współczynnika korelacji ρ jest wielkość r zwana współczynnikiem korelacji Pearsona i określona wzorem

$$r = \frac{\sum_i \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum_i \left(x_i - \bar{x}\right)^2 \sum_i \left(y_i - \bar{y}\right)^2}} \tag{8.9}$$

lub wzorem dogodniejszym do obliczeń:

$$r = \frac{\sum_i x_i y_i - n\,\bar{x}\,\bar{y}}{\sqrt{\left(\sum_i x_i^2 - n\,\bar{x}^2\right)\left(\sum_i y_i^2 - n\,\bar{y}^2\right)}} \tag{8.10}$$

Estymatory współczynników kierunkowych prostych regresji: y względem x (b_{yx}) i x względem y (b_{xy}) oraz współczynnik korelacji Pearsona powiązane są zależnościami:

$$b_{yx} = r\frac{s_y}{s_x} \qquad b_{xy} = r\frac{s_x}{s_y} \tag{8.11}$$

$$r^2 = b_{yx}\,b_{xy} \tag{8.12}$$

gdzie s_y i s_x są estymatorami odchyleń standardowych zmiennych y i x.

8.2 Estymacja przedziałowa parametrów prostej regresji, wnioskowanie o istotności związku prostoliniowego

W literaturze podaje się następujące wzory na oszacowania wariancji statystyk a i b — czyli parametrów estymowanej prostej regresji (8.2):

$$\sigma^2(b) = \frac{s_0^2}{\sum_i \left(x_i - \bar{x}\right)^2} \tag{8.13}$$

$$\sigma^2(a) = s_0^2\left[\frac{1}{n} + \frac{\bar{x}}{\sum_i \left(x_i - \bar{x}\right)^2}\right] \tag{8.14}$$

gdzie: s_0^2 — jest oszacowaniem wariancji resztowej odchyleń od funkcji regresji

$$s_0^2 = \frac{\sum_i \left(y_i - \bar{y}\right)^2 (1-r)^2}{n-2} = \frac{\sum_i \left(y_i - Y_i\right)^2}{n-2} \tag{8.15}$$

Wzory (8.13) i (8.14) mogą być użyte do weryfikacji hipotez dotyczących a i b. Weryfikację przeprowadza się wykorzystując test t-Studenta przy $n - 2$ stopniach swobody. Dla przykładu hipotezę zerową $H_0 : B = 0$ mówiącą, że prosta regresji jest równoległa do osi x (czyli średnia wartość y nie zmienia się ze zmianą x) testujemy porównując statystykę t

$$t = \frac{b}{\sqrt{\sigma^2(b)}} \tag{8.16}$$

z wartością krytyczną rozkładu t dla $n - 2$ stopni swobody. Jeżeli więc

$$|t| \geq {}_{0,05}\, t_{(n-2)}$$

to hipotezę zerową odrzucamy na poziomie istotności 0,05, przyjmując hipotezę alternatywną $H_1 : B \neq 0$.

Częściej stosowanym (choć równoważnym poprzedniemu — por. zależność (8.11)) testem jest weryfikacja hipotezy zerowej $H_0 : \rho = 0$ mówiącej, że współczynnik korelacji jest równy zeru, czyli że zmienne x i y nie są skorelowane. Obliczamy wówczas statystykę t

$$t = \frac{r}{\sqrt{\sigma^2(r)}} \tag{8.17}$$

gdzie:

$$\sigma^2(r) = \frac{1-r^2}{n-2} \tag{8.18}$$

i porównujemy jej wartość, podobnie jak w poprzednim teście, z wartością krytyczną rozkładu t dla $n - 2$ stopni swobody przy zadanym poziomie istotności. Często zamiast stosować tę procedurę można skorzystać ze specjalnych tablic wartości krytycznych współczynnika korelacji.

Przedziały ufności dla A i B określamy w znany sposób. Przykładowo 95%-owy przedział ufności dla B można określić jako

$$b \pm {}_{0,05}\, t_{(n-2)}\, \sigma(B) \tag{8.19}$$

Procedura postępowania przy wyznaczaniu przedziału ufności dla ρ jest nieco inna. Zastępujemy wartość r statystyką Z

$$Z = \frac{1}{2} \ln \frac{1+r}{1-r} \tag{8.20}$$

Rozkład z próby wielkości Z jest w przybliżeniu normalny nawet dla małych prób, a także dla prób o rozkładach odbiegających nieco od normalnego. Błąd standardowy Z dany jest wzorem:

$$\sigma(Z) = \frac{1}{\sqrt{n-3}} \tag{8.21}$$

Granice przedziału ufności dla Z przy współczynniku ufności $1 - \alpha$ uzyskujemy z formuły

$$Z \pm {}_{\alpha} u \, \sigma(Z) \tag{8.22}$$

Stosując przekształcenie odwrotne do (8.20)

$$r = \frac{e^{2Z} - 1}{e^{2Z} + 1} \tag{8.23}$$

uzyskujemy granice ufności dla ρ. Przedział tak uzyskany będzie niesymetryczny czego powodem jest niesymetria rozkładu współczynnika korelacji r dla $\rho \neq 0$. Ta właśnie niesymetria powoduje, że dla testowania hipotezy $H_0 : \rho = \rho_0$, gdzie $\rho_0 \neq 0$, nie można stosować testu (8.17) (8.18), lecz należy wykorzystać statystykę u

$$u = \frac{1}{2} \left(\ln \frac{1+r}{1-r} - \ln \frac{1+\rho_0}{1-\rho_0} - \frac{\rho_0}{n-1} \right) \sqrt{n-3} \tag{8.24}$$

mającą w przybliżeniu rozkład normalny standaryzowany. Hipotezę H_0 odrzucamy, gdy

$$|u| \geq {}_{\alpha} u$$

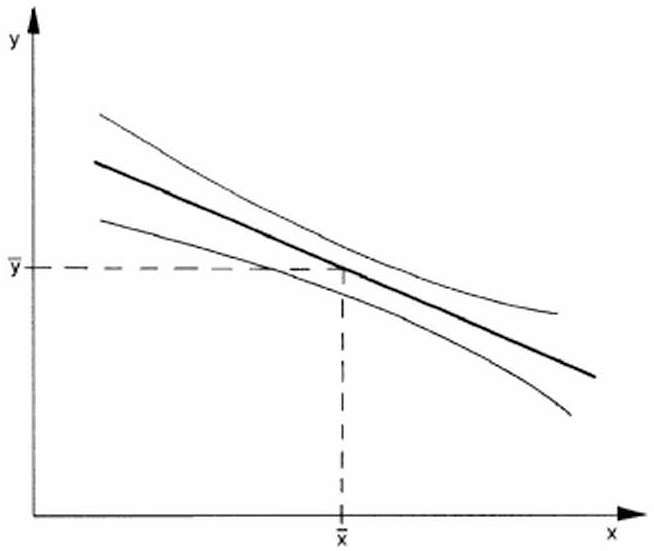
Innym zagadnieniem związanym z regresją jest predykcja. Chodzi o znalezienie przewidywanej wartości Y zmiennej y odpowiadającej danemu x_0. Najlepszym oszacowaniem wartości przewidywanej jest wartość Y_0 wynikająca z prostej regresji, czyli

$$Y_0 = a + b \cdot x_0 \tag{8.25}$$

Granice błędu dla Y_0 związane z losowością samego y oraz niedokładnością określenia parametrów prostej regresji wyrażają się wzorem:

$$Y_0 = \alpha\, t_{(n-2)} \cdot s_0 \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})2}{\sum_i (x_i - \bar{x})2}} \tag{8.26}$$

Wzór (8.26) pozwala na określenie obszaru, w którym z prawdopodobieństwem $1 - \alpha$ znajdą się przewidywane wartości Y dla różnych wartości x_0 zmiennej x. Ilustruje to rysunek 8.3.



Rys. 8.3 95-procentowe granice ufności dla przewidywanych wartości Y przy określonych wartościach x_0 zmiennej x.

Przykład 8.1 (zmodyfikowany według [Armitage])

Przez dwa tygodnie karmiono szczury dietą ubogą w witaminę D dla wywołania krzywicy. Następnie przez dalsze dwa tygodnie podawano szczurom preparat zawierający witaminę D. Po upływie tego czasu określano stopień wyleczenia przez radiografię prawego kolana każdego zwierzęcia doświadczalnego. Porównywano analizowane zdjęcie radiologiczne ze standardowym zestawem fotografii opatrzonych numerami od 0 do 12 (stopniowanie w kierunku wzrastającego wyleczenia). Każdą dawkę preparatu podawano grupie kilku szczurów i późniejsze zdjęcia analizowało kilku radiologów. Zbadać regresję liniową między logarytmem dawki preparatu a średnim efektem dla każdej dawki.

Wyniki badań zawarto w tabeli 8.1. Tamże zamieszczono pewne pośrednie wyniki obliczeń.

Tabela 8.1

Wyniki badań efektu terapeutycznego preparatu przeciwkrzywicznego

	Obserwacje								Sumy
dawka	2,5	5	10	20	40	80	160	320	
x_i	0,398	0,699	1,000	1,301	1,602	1,903	2,204	2,505	11,612
y_i	0,250	1,0833	1,6667	2,8333	3,5833	4,3333	4,9167	5,5555	24,0000
$x_i \cdot y_i$	0,0995	0,7572	1,6667	3,3862	5,7404	8,2463	10,8363	13,3600	44,3927
x_i^2	0,1584	0,4886	1,0000	1,6926	2,5664	3,6214	4,8576	6,2750	20,6601
y_i^2	0,0625	1,1736	2,7778	8,0278	12,8402	18,7778	24,1736	28,4444	96,2778
x_i — logarytm dawki					y_i — średni efekt				

Mamy:

$$\bar{x} = 1{,}4515 \qquad \bar{y} = 3{,}00 \qquad n = 8$$

$$\sum x_i y_i - n\bar{x}\bar{y} = 9{,}5567$$

$$\sum x_i^2 - n\bar{x}^2 = 3{,}8052$$

$$\sum y_i^2 - n\bar{y}^2 = 24{,}2778$$

Stąd:

$$a = 2{,}5115$$

$$b = -0{,}6454$$

$$r = 0{,}9943$$

Przeprowadzimy test współczynnika korelacji sprawdzając, czy można uznać, że jest on istotnie różny od zera.

$$t = r\sqrt{\frac{n-2}{1-r^2}} = 22{,}8296$$

Ponieważ

$$|t| > {}_{0,05}t_{(6)} = 2{,}447$$

uznajemy, że korelacja między zmiennymi x i y jest istotna. Chcąc wyznaczyć granice przedziału ufności dla współczynnika korelacji obliczamy:

$$Z = 2{,}9281$$

$$\sigma(Z) = \frac{1}{\sqrt{5}} = 0{,}4472$$

Dla współczynnika ufności równego 0,95 mamy

$${}_{0,05}u = 1{,}96$$

Przedział ufności dla Z określimy jako

$$2{,}9281 \pm 0{,}8765$$

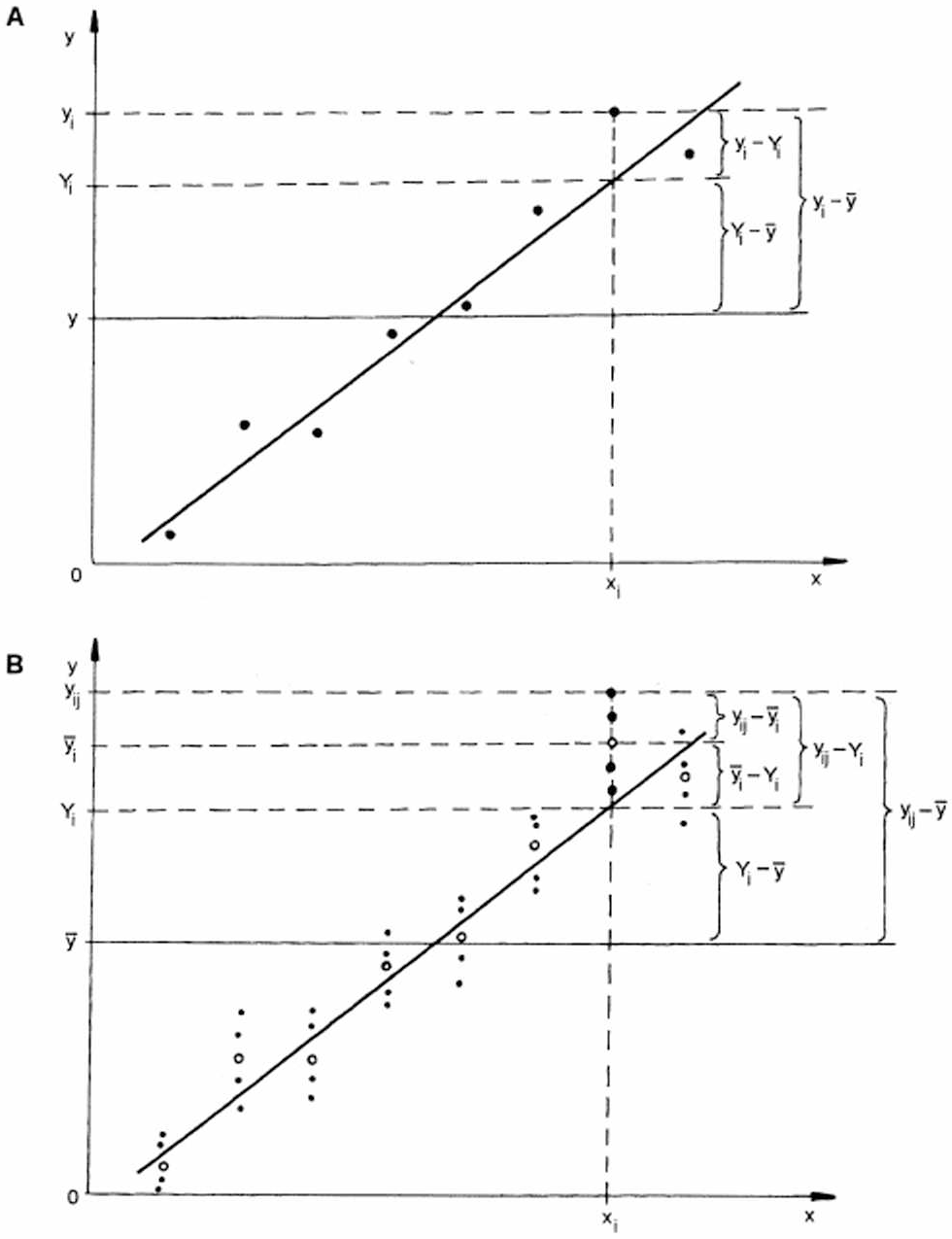
czyli jego granice będą równe 2,0516 i 3,8046. Stosując do nich przekształcenie (8.23) otrzymamy ostatecznie, że 95%-owe granice przedziału ufności dla współczynnika korelacji wynoszą 0,9674 i 0,9986. Niesymetria tego przedziału jest oczywista, zważywszy, że $r = 0{,}9943$.

8.3 Zastosowanie analizy wariancji do problemów związanych z regresją. Test na liniowość

W podrozdziale tym pokażemy, w jaki sposób można wykorzystać metodę analizy wariancji do badania problemów związanych z regresją liniową. Niech prosta regresji uzyskana przy pomocy znanych wzorów będzie określona równaniem:

$$Y = a + b \cdot x \tag{8.27}$$

Dla każdej wartości x_i dysponujemy wartością obserwowaną y_i i wartością teoretyczną Y_i wynikającą z równania regresji (8.27). Odchylenie wartości obserwowanej y_i od wartości średniej $\bar{y}$ można rozbić na dwa składniki: odchylenie wartości obserwowanej y_i od teoretycznej Y_i oraz odchylenie wartości teoretycznej Y_i od średniej $\bar{y}$ (por. rys. 8.4.A)



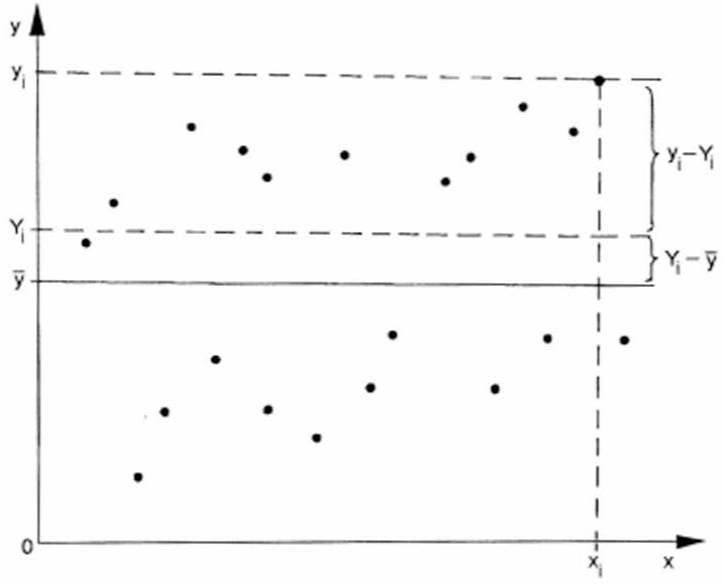
Rys. 8.4 Podział odchyleń wartości obserwacji od średniej:
A — pojedyncze obserwacje,
B — serie obserwacji.

$$y_i - \bar{y} = \left(y_i - Y_i\right) + \left(Y_i - \bar{y}\right)$$

Można wykazać, że podobna zależność zachodzi dla sum kwadratów odchyleń:

$$\sum_i \left(y_i - \bar{y}\right)^2 = \sum_i \left(y_i - Y_i\right)^2 + \sum_i \left(Y_i - \bar{y}\right)^2 \tag{8.28}$$

Na rysunku 8.4.A prawie cała ogólna suma kwadratów odchyleń od średniej $\sum (y_i - \bar{y})^2$ wyjaśniona jest przez sumę kwadratów odchyleń wartości teoretycznych od średniej $\sum (Y_i - \bar{y})^2$. Jednak nie zawsze tak jest. Rysunek 8.5 przedstawia przypadek przeciwny. Tutaj nachylenie prostej regresji jest niewielkie, rozrzut wokół linii regresji duży i to powoduje, że prawie cała ogólna suma kwadratów odchyleń od średniej jest objaśniona przez sumę kwadratów odchyleń od linii regresji, tzn. odchyleń wartości obserwowanych od wartości teoretycznych $\sum (y_i - Y_i)^2$. Takie rozumowanie nasuwa skojarzenia z rozważaniami stojącymi u podstaw metody analizy wariancji. Dzieląc omawiane sumy kwadratów przez odpowiednie liczby stopni swobody i testując iloraz średniego kwadratu odchyleń wartości teoretycznych od średniej przez średni kwadrat odchyleń od prostej regresji, weryfikujemy hipotezę zerową $H_0 : B = 0$, mówiącą o zerowym nachyleniu prostej regresji, czyli o braku związku liniowego pomiędzy zmiennymi x i y. Test taki będzie całkowicie równoważny testowi t przedstawionemu w podrozdziale 8.2 (por. wzór (8.16)).



Rys. 8.5 Podział odchyleń wartości obserwacji od średniej w analizie regresji.

Zalety stosowania analizy wariancji do problemów związanych z regresją widać dopiero wówczas, gdy dysponujemy kilkoma wartościami obserwacji zmiennej y dla pewnego $x = x_i$. Każdą taką grupę obserwacji dla danego x_i nazywamy serią. Niech liczebność i-tej serii będzie oznaczona przez n_i, całkowita liczba obserwacji przez N, zaś liczba serii przez k. J-tą obserwację w i-tej serii oznaczymy przez y_{ij}. Pozostałe oznaczenia pokazano w tabeli 8.2.

Przeanalizujemy teraz sytuację przedstawioną na rysunku 8.4.B. Odchylenie wartości obserwacji od wartości teoretycznej rozbijemy na dwa składniki: odchylenie wartości obserwacji y_{ij} od średniej serii $\bar{y}_i$ oraz odchylenie średniej serii od wartości teoretycznej Y_i. Całkowitą sumę kwadratów odchyleń od wartości średniej $\bar{y}$ można zatem rozbić na trzy składniki

Tabela 8.2

Oznaczenia stosowane w teście na liniowość funkcji regresji wykorzystującym analizę wariancji

x_i	x_1	x_2 . . . x_i . . . x_k	Ogółem
y_{ij}	y_{11} y_{12} : y_{1n_1}	y_{21} y_{i1} y_{k1} y_{22} y_{i2} y_{k2} : : : y_{2n_2} y_{in_i} y_{kn_k}	
$T_i = \sum_{j=1}^{n_i} y_{ij}$	T_1	T_2 T_i T_k	$T = \sum_{i=1}^{k_i} T_i$
$s_i = \sum_{j=1}^{n_i} y_{ij}^2$	s_1	s_2 s_i s_k	$s = \sum_{i=1}^{k_i} s_i$
$\bar{y}_i = \frac{T_i}{n_i}$	$\bar{y}_1$	$\bar{y}_2$ $\bar{y}_i$ $\bar{y}_k$	$\bar{y} = \frac{T}{N}$
n_i	n_1	n_2 n_i n_k	$N = \sum_{i=1}^{k_i} n_i$

$$\sum_{i,j}\left(y_{ij}-\bar{y}\right)^2=\sum_{i,j}\left(y_{ij}-\bar{y}_i\right)^2+\sum_{i,j}\left(\bar{y}_i-Y_i\right)^2+\sum_{i,j}\left(Y_i-\bar{y}\right)^2 \tag{8.29}$$

Suma kwadratów odchyleń wartości obserwacji od średnich serii (*SKR*)

$$SKR=\sum_{i,j}\left(y_{ij}-\bar{y}_i\right)^2=S-\sum_{i=1}^{k}\frac{T_i^2}{n_i} \tag{8.30}$$

jest miarą zmienności resztowej, czyli wewnętrznej zmienności losowej materiału obserwacyjnego. *SKR* podzielona przez liczbę stopni swobody równą $N - k$ służy jako wartość odniesienia w teście analizy wariancji dla pozostałych składników ogólnej sumy kwadratów. Suma kwadratów odchyleń wartości teoretycznych od średniej *SKB*

$$SKB=\sum_{i,j}\left(Y_i-\bar{y}\right)^2=\frac{\left(\sum_{i=1}^{k}x_i T_i-\frac{T_i\cdot\sum_{i=1}^{k}n_i x_i}{N}\right)^2}{\sum_{i=1}^{k}n_i x_i^2-\frac{\left(\sum_{i=1}^{k}n_i x_i\right)^2}{N}} \tag{8.31}$$

jest wielkością obdarzoną jednym stopniem swobody. Porównana ze średnim kwadratem resztowym s_0^2

$$s_0^2=\frac{SKR}{N-k}$$

(por. tabela 8.3) służy do weryfikacji hipotezy zerowej $H_0 : B = 0$ (stosunek F_1). Sumę kwadratów odchyleń średnich serii od prostej regresji *SKL* obliczamy zgodnie ze wzorem:

$$SKL=\sum_{i,j}\left(\bar{y}_i-Y_i\right)^2=SK-SKR-SKB$$

przy czym ogólna suma kwadratów *SK* dana jest zależnością:

$$SK=\sum_{i,j}\left(y_{ij}-\bar{y}\right)^2=S-\frac{T^2}{N} \tag{8.32}$$

Tabela 8.3

Tablica analizy wariancji dla regresji liniowej z testem na liniowość

Źródło zmienności	Suma kwadratów	Liczba stopni swobody	Średni kwadrat	Stosunek wariancji
Odchylenia wartości teoretycznej od średniej	SKB	1	$s_1^2 = \frac{SKB}{1}$	$F_1 = \frac{s_1^2}{s_0^2}$
Odchylenia średnich od prostej regresji	SKL	$k - 2$	$s_2^2 = \frac{SKL}{k-2}$	$F_2 = \frac{s_2^2}{s_0^2}$
Reszta wewnątrz serii	SKR	$N - k$	$s_0^2 = \frac{SKR}{N-k}$	
Całkowita	SK	$N - 1$		

Suma kwadratów SKL podzielona przez liczbę stopni swobody równą $k - 2$ i porównana ze średnim kwadratem resztowym s_0^2 (tabela 8.3) służy do weryfikacji hipotezy o liniowości funkcji regresji. Jeżeli stosunek wariancji F_2 daje wartość większą niż wartość krytyczna przy $k - 2$ i $N - k$ stopniach swobody, to hipotezę o prostoliniowości funkcji regresji należy odrzucić. Trzeba wówczas zastanowić się, czy nie uzyskalibyśmy lepszych rezultatów dopasowując do naszych danych jakąś nieprostoliniową postać funkcji regresji.

Przykład 8.2

Wróćmy do badań przedstawionych w przykładzie 8.1. Tam dopasowywaliśmy linię regresji do zależności wzajemnej logarytmu dawki leku (x_i) i średniego efektu terapeutycznego (y_i). Tutaj zamiast efektu średniego uwzględnimy, że każdą dawkę leku (x_i) podawano sześciu szczurom. Wobec tego dysponowano sześcioma radiologicznymi ocenami efektu leczenia dla każdego x_i — długość każdej analizowanej serii wynosiła 6.

Tabela 8.4

Wyniki badań efektu terapeutycznego preparatu przeciwkrzywicznego (preparat 1)

dawka	2,5	5	10	20	40	
x_i – logarytm dawki	0,398	0,699	1,000	1,301	1,602	$\bar{x} = 1$
y_{ij} – oceny radiologiczne efektu	0	1,0	1,5	3,0	6,5	
	0	1,5	1,0	3,0	3,5	
	0	1,5	2,0	5,5	4,5	
	0	1,0	3,5	2,5	3,5	
	0	1,0	2,0	1,0	3,5	
	0,5	0,5	0	2,0	3,0	
n_i	6	6	6	6	6	$N = 30$
T_i	0,5	6,5	10,0	17,0	24,5	$T = 58,5$
$\bar{y}_i$	0,0833	1,0833	1,6667	2,8333	4,0833	$\bar{y} = 1,95$
S_i	0,25	7,75	23,5	59,5	108,25	$S = 199,25$

Dane oraz wyniki niektórych obliczeń pośrednich przedstawiono w tabeli 8.4 (uwzględniono 5 różnych dawek preparatu). Estymowane parametry prostej regresji wynoszą w tym przypadku:

$a = -1,2892$
$b = 3,2392$

Przedstawioną powyżej metodą zbadano istotność współczynnika regresji B efektu terapii w zależności od logarytmu dawki preparatu oraz liniowość funkcji regresji. Wyniki pokazuje tabela 8.5. Widać, że odchylenia funkcji regresji od prostej są nieistotne, przyjmujemy więc (niezbyt formalnie), że regresja jest rzeczywiście prostoliniowa. Nachylenie linii regresji, jak to widać z testu stosunku F_1, jest wysoce istotne, tzn. musimy odrzucić testowaną hipotezę, że $B = 0$.

Tabela 8.5

Tablica analizy wariancji dla regresji liniowej z testem na liniowość (dla danych z tablicy 8.4)

Źródło zmienności	Suma kwadratów	Liczba stopni swobody	Średni kwadrat	Stosunek wariancji	Istotność
Odchylenia wartości teoretycznej od średniej	57,0375	1	57,0375	52,248	$P < 0{,}005$
Odchylenia średniego efektu dawki od prostej	0,8458	3	0,2819	0,258	nieistotne
Reszta wewnątrz dawki	27,2917	25	1,0917		
Ogółem	85,175	29			

8.4. Regresja w grupach

Często dysponujemy k grupami obserwacji dwucechowych, to znaczy posiadamy w każdej l-tej grupie obserwacji ($l = 1,2,...,k$) n_l par wartości (x_{il}, y_{il}). Odpowiednie wartości średnie oznaczymy $\overline{x}_l$ oraz $\overline{y}_l$. Spodziewamy się, że obserwacje w każdej grupie układają się wokół pewnej prostej regresji. Chcemy porównać nachylenia tych prostych regresji, a mówiąc dokładnie sprawdzić, czy można uważać poszczególne linie regresji w grupach za równoległe. Linie regresji dopasowane indywidualnie do obserwacji w każdej grupie bardzo rzadko będą posiadały identycznie takie samo nachylenie. Będziemy więc sprawdzać, czy różnice między nachyleniami mogą być wytłumaczone tylko zmiennością losową i proste te mogą być uważane za równoległe, czy też nachylenia są istotnie różne. Jeżeli już stwierdzimy, że nachylenia prostych regresji mogą być uznane za jednakowe, to będziemy chcieli jeszcze rozstrzygnąć czy równoległe proste regresji pokrywają się, czy też można uważać ich położenia za istotnie różne.

8.4.1 Test rownoległości prostych regresji dla dwóch grup

Jeżeli mamy do czynienia z dwoma tylko grupami obserwacji ($k = 2$) i możemy założyć równość wariancji resztowych w obu grupach, to estymatorem wspólnej wariancji resztowej (średniego kwadratu odchyleń od prostych regresji będzie

$$s^2 = \frac{\sum_{i=1}^{n_1}\left(y_{i1} - Y_{i1}\right)^2 + \sum_{i=1}^{n_2}\left(y_{i2} - Y_{i2}\right)^2}{n_1 + n_2 - 4} \tag{8.33}$$

Niech b_1 i b_2 będą oszacowaniami współczynników regresji każdej z grup. Wtedy wariancję różnicy $b_1 - b_2$ można oszacować jako:

$$\sigma^2\left(b_1 - b_2\right) = s^2 \cdot \left[\frac{1}{\sum_{i=1}^{n_1}\left(x_{i1} - \bar{x}_1\right)^2} + \frac{1}{\sum_{i=1}^{n_2}\left(x_{i2} - \bar{x}_2\right)^2}\right] \tag{8.34}$$

Weryfikując hipotezę zerową $H_0 : B_1 = B_2$ o równoległości obu prostych regresji testujemy różnicę $b_1 - b_2$ obliczając statystykę t

$$t = \frac{b_1 - b_2}{\sqrt{\sigma^2\left(b_1 - b_2\right)}} \tag{8.35}$$

i porównując ją z wartością krytyczną rozkładu t przy $n_1 + n_2 - 4$ stopniach swobody i ustalonym poziomie istotności α. Gdy

$$|t| \geq {}_{\alpha}t_{(n_1 + n_2 - 4)}$$

hipotezę o równoległości prostych regresji należy odrzucić. Obliczenia według wzoru (8.33) są utrudnione, gdyż wymagają wyliczenia wartości teoretycznych. Można je uprościć stosując wzór

$$\sum_{i=1}^{n_1}\left(y_{i1} - Y_{i1}\right)^2 = \sum_{i=1}^{n_1}\left(y_{i1} - \bar{y}_i\right)^2 - \frac{\left[\sum_{i=1}^{n_1}\left(x_{i1} - \bar{x}_1\right)\left(y_{i1} - \bar{y}_1\right)^2\right]^2}{\sum_{i=1}^{n_1}\left(x_{i1} - \bar{x}_1\right)^2} \tag{8.36}$$

i zastępując odpowiednie sumy kwadratów lub iloczynów wyrażeniami:

$$\sum_{i=1}^{n_1}\left(x_{i1}-\overline{x}_1\right)^2=\sum_{i=1}^{n_1}x_{i1}^2-\frac{\left(\sum_{i=1}^{n_1}x_{i1}\right)^2}{n_1} \tag{8.37}$$

$$\sum_{i=1}^{n_1}\left(y_{i1}-\overline{y}_1\right)^2=\sum_{i=1}^{n_1}y_{i1}^2-\frac{\left(\sum_{i=1}^{n_1}y_{i1}\right)^2}{n_1} \tag{8.38}$$

$$\sum_{i=1}^{n_1}\left(x_{i1}-\overline{x}_1\right)\left(y_{i1}-\overline{y}_1\right)=\sum_{i=1}^{n_1}x_{i1}\,y_{i1}-\frac{\sum_{i=1}^{n_1}x_{i1}\sum_{i=1}^{n_1}y_{i1}}{n_1} \tag{8.39}$$

Podobnie należy postąpić z tymi fragmentami wzoru (8.33), które odnoszą się do grupy drugiej. Z powyższych wzorów, odpowiednio adaptowanych, należy również korzystać w dalszych obliczeniach.

Jeżeli nie ma podstaw do odrzucenia hipotezy o równoległości prostych regresji, wyznaczamy wartość b wspólnego współczynnika nachylenia obu prostych regresji w grupach jako:

$$b=\frac{\sum_{i=1}^{n_1}\left(x_{i1}-\overline{x}_1\right)\left(y_{i1}-\overline{y}_1\right)+\sum_{i=1}^{n_2}\left(x_{i2}-\overline{x}_2\right)\left(y_{i2}-\overline{y}_2\right)}{\sum_{i=1}^{n_1}\left(x_{i1}-\overline{x}_1\right)^2+\sum_{i=1}^{n_2}\left(x_{i2}-\overline{x}_2\right)^2} \tag{8.40}$$

oraz oszacowanie jego wariancji:

$$\sigma^2(b)=\frac{s^2}{\sum_{i=1}^{n_1}\left(x_{i1}-\overline{x}_1\right)^2+\sum_{i=1}^{n_2}\left(x_{i2}-\overline{x}_2\right)^2} \tag{8.41}$$

Granice ufności dla b można oszacować jako

$$b\pm{}_{\alpha}t_{(n_1+n_2-4)}\sqrt{\sigma^2(b)} \tag{8.42}$$

Przykład 8.3

Opisane w przykładach 8.1 i 8.2 badania przeprowadzono dla trzech preparatów przeciwkrzywicznych, zawsze oceniając efekt terapeutyczny przez uśrednienie kilku ocen zdjęć radiologicznych. Dane zamieszczono w tabeli 8.6, zaś pośrednie wyniki obliczeń zawiera tabela 8.7.

Tabela 8.6

Wyniki badań efektu terapeutycznego trzech pacjentów

Preparat 1

dawka	2,5	5	10	20	40
logarytm dawki	0,398	0,699	1,000	1,301	1,602
oceny radiologiczne efektu	0 0 0 0 0 0,5	1,0 1,5 1,5 1,0 1,0 0,5	1,5 1,0 2,0 3,5 2,0 0	3,0 3,0 5,5 2,5 1,0 2,0	6,5 3,5 4,5 3,5 3,5 3,0

Preparat 2

dawka	2,5	5	10	20
logarytm dawki	0,398	0,699	1,000	1,301
oceny radiologiczne efektu	3,25 3,0 1,75 2,0 0,5 1,5	2,50 2,75 2,25 2,25 3,75	3,75 5,25 6,0 5,5 2,25 3,50	5,0 7,0 3,0 3,0 6,0

Preparat 3

dawka	3,5	7	14	28
logarytm dawki	0,544	0,845	1,146	1,447
oceny radiologiczne efektu	1,0 2,5 0 3,0 2,0	1,5 5,0 2,0 3,5 2,0	2,0 3,0 4,5 6,0 3,5	3,5 6,5 2,0 5,0 5,0

Tabela 8.7

Zebrane pośrednie wyniki obliczeń dotyczących danych z tabeli 8.6

l	$\sum_i x_{il}$	$\sum_i y_{il}$	$\sum_i x_{il} y_{il}$	$\sum_i x_{il}^2$	$\sum_i y_{il}^2$
1	30,0	58,5	76,1085	35,4361	199,25
2	18,388	75,75	71,6865	17,8564	320,3125
3	19,91	64,0	70,4845	22,0854	260,0
Ogółem	68,298	198,25	218,2795	75,3779	779,5625

l	$(Sx^2)_l$	$(Sy^2)_l$	$(Sxy)_l$	$(Sy^2)_l - \frac{(Sxy)_l}{(Sx^2)_l}$
1	5,4361	85,175	17,6085	28,1375
2	2,4874	59,4915	8,3733	31,3049
3	2,2650	55,2	6,7725	34,9500
Ogółem	10,1885	199,8665	32,7543	94,3924

l	n_l	$\bar{x}_l$	$\bar{y}_l$	b_l	a_l
1	30	1,0	1,95	3,2392	–1,2892
2	22	0,8358	3,4432	3,3663	0,6296
3	20	0,9955	3,2	2,9900	0,2234
Ogółem	72			(3,2148)	

Tutaj będziemy porównywać nachylenia prostych regresji dla preparatu 1 i dla preparatu 2 wykorzystując sposób opisany w podpunkcie 8.4.1. Estymator s^2 wspólnej wariancji resztowej obliczony na podstawie (8.33) przy wykorzystaniu (8.37) ÷ (8.39) wynosi:

$$s^2 = \frac{28{,}1375 + 31{,}3049}{30 + 22 - 4}$$

zaś wariancja różnicy współczynników regresji $b_1 - b_2$ jest równa

$$\sigma^2 (b_1 - b^2) = 1{,}2384 \left(\frac{1}{5{,}4361} + \frac{1}{2{,}4874} \right) = 0{,}7257$$

Obliczając statystykę t uzyskamy:

$$t = \frac{3{,}2392 - 3{,}3663}{\sqrt{0{,}7257}} = -0{,}1492$$

co jest oczywiście mniejsze co do modułu od wartości krytycznej ${}_{0{,}05}t_{(48)} = 2{,}013$. Różnica między nachyleniami prostych regresji nie jest więc istotna. Wspólny współczynnik nachylenia obu równoległych prostych regresji będzie więc równy

$$b = \frac{17{,}6085 + 8{,}3733}{5{,}4361 + 2{,}4874} = 3{,}2791$$

jego wariancja będzie wynosić

$$\sigma^2 (b) = \frac{1{,}2384}{5{,}4361 + 2{,}4874} = 0{,}1563$$

a 95%-owe granice przedziału ufności β będą równe

$$3{,}2791 \pm 2{,}013 \sqrt{0{,}7257} = 4{,}0738 \quad \text{oraz} \quad 2{,}4844$$

8.4.2 Test równoległości prostych regresji dla kilku grup

W przypadku potrzeby porównywania współczynników regresji dla k grup ($k > 2$) będziemy używać metody analizy wariancji. Dla uproszczenia zapisu przyjmiemy oznaczenia:

$$(Sx^2)_l = \sum_i \left(x_{il} - \overline{x}_l\right)^2 \tag{8.43}$$

$$(Sy^2)_l = \sum_i \left(y_{il} - \overline{y}_l\right)^2 \tag{8.44}$$

$$(Sxy)_l = \sum_i \left(x_{il} - \overline{x}_l\right)\left(y_{il} - \overline{y}\right) \tag{8.45}$$

Do praktycznych obliczeń wielkości (Sx^2), (Sy^2) i (Sxy) warto wykorzystywać wzory (8.37) ÷ (8.39).

Wspólny współczynnik nachylenia k równoległych prostych regresji oblicza się jako

$$b = \frac{\sum_l (Sxy)_l}{\sum_l (Sx^2)_l} \tag{8.46}$$

Dla potrzeb testowania hipotezy zerowej o równości współczynników regresji (równoległości prostych regresji) w k grupach rozbijamy sumę kwadratów odchyleń wartości y_{il} od średnich grupowych $\overline{y}_l$ $(SKWG)$

$$SKWG = \sum_{i,l}\left(y_{il} - \overline{y}_l\right)^2 = \sum_{i,l}\left(y_{il} - Y_{il}\right)^2 + \sum_{i,l}\left(Y_{il} - Y_{il}^{(c)}\right)^2 + \sum_{i,l}\left(Y_{il}^{(c)} - \overline{y}_l\right)^2 \tag{8.47}$$

na trzy składniki. Pierwszy z nich

$$SKRW = \sum_{i,l}\left(y_{il} - Y_{il}\right)^2 = \sum_l (Sx^2)_l - \sum_l \frac{(Sxy)_l^2}{(Sx^2)_l} \tag{8.48}$$

jest resztową sumą kwadratów odchyleń obserwacji od poszczególnych prostych regresji, indywidualnych dla każdej grupy

$$Y_{il} = a_l + b_l \cdot x_{il} \tag{8.49}$$

Drugi składnik

$$SKMW = \sum_{i,l} \left(Y_{il} - Y_{il}^{(c)} \right)^2 = \sum_{l} \frac{(Sxy)_l^2}{(Sx^2)_1} - \frac{\left[\sum_{l} (Sxy)_l \right]^2}{\sum_{l} (Sx^2)_l} \tag{8.50}$$

to suma kwadratów odchyleń wartości wskazywanych przez poszczególne proste regresji (8.49) od wartości wynikających z równoległych prostych regresji o jednakowym współczynniku nachylenia b

$$Y_{il}^{(c)} = a_l^{(c)} + b\, x_{il} \tag{8.51}$$

Ostatnim składnikiem

$$SKWW = \sum_{i,l} \left(Y_{il}{}^{(c)} - \bar{y}_l \right)^2 = \frac{\left[\sum_{l} (Sxy)_l \right]^2}{\sum_{l} (Sx^2)_l} \tag{8.52}$$

jest suma kwadratów odchyleń wartości wskazywanych przez równoległe proste regresji od średnich w grupach.

Tablicę analizy wariancji w tym przypadku przedstawia tabela 8.8. Przyjęto tam oznaczenie sumarycznej liczebności obserwacji literą N

$$N = \sum_{l=1}^{k} n_l$$

Test stosunku F_1 przy 1 i $N - 2k$ stopniach swobody pozwala stwierdzić istotność ogólnego nachylenia prostych równoległych. Nas będzie szczególnie interesował stosunek F_2, którego przetestowanie przy $k - 1$ i $N - 2k$ stopniach swobody pozwala odrzucić hipotezę o równości współczynników nachylenia poszczególnych prostych regresji w grupach.

Tablica analizy wariancji dla potrzeb testu równoległości prostych regresji w kilku grupach

Źródło zmienności	Suma kwadratów	Liczba stopni swobody	Średni kwadrat	Stosunek wariancji
Odchylenia od prostych o wspólnym współczynniku nachylenia	$SKWW$	1	$s_1^2 = \frac{SKWW}{1}$	$F_1 = \frac{s_1^2}{s_0^2}$
Różnice między współczynikami nachylenia poszczególnych prostych	$SKMW$	$k - 1$	$s_2^2 = \frac{SKMW}{k-1}$	$F_2 = \frac{s_2^2}{s_0^2}$
Reszta odchyleń od poszczególnych linii regresji	$SKRW$	$N - 2k$	$s_0^2 = \frac{SKRW}{N-2k}$	
Całkowita zmienność wewnątrz grup	$SKWG$	$N - k$		

Przykład 8.4

Chcemy porównać współczynniki nachylenia prostych regresji otrzymanych podczas badań efektu terapeutycznego trzech preparatów przeciwkrzywicznych (tabele 8.6, 8.7, patrz także przykład 8.3).

Wspólny współczynnik kierunkowy trzech równoległych prostych regresji będzie wynosił:

$$b = \frac{17{,}6085 + 8{,}3733 + 6{,}7725}{5{,}4361 + 2{,}4874 + 2{,}2650} = 3{,}2148$$

Tablicę analizy wariancji pokazuje tabela 8.9.

Tabela 8.9

Tablica analizy wariancji przedstawiająca wyniki badania równoległości prostych regresji dla danych z tabeli 8.6

Źródło zmienności	Suma kwadratów	Liczba stopni swobody	Średni kwadrat	Stosunek wariancji	Istotność
Ogólne nachylenie równoległych prostych	105,2994	1	105,2994	73,6263	$P < 0{,}005$
Różnice między nachyleniami poszczególnych prostych	0,1747	2	0,0874	0,0611	nieistotne
Reszty wewnątrz grupy	94,3924	66	1,4302		
Ogółem wewnątrz grupy	199,8665	69			

Ogólne nachylenie równoległych prostych regresji o współczynniku kierunkowym b jest wysoce istotne

$$F_1 >> {}_{0,005}F^{(1)}_{(66)} = 8{,}459$$

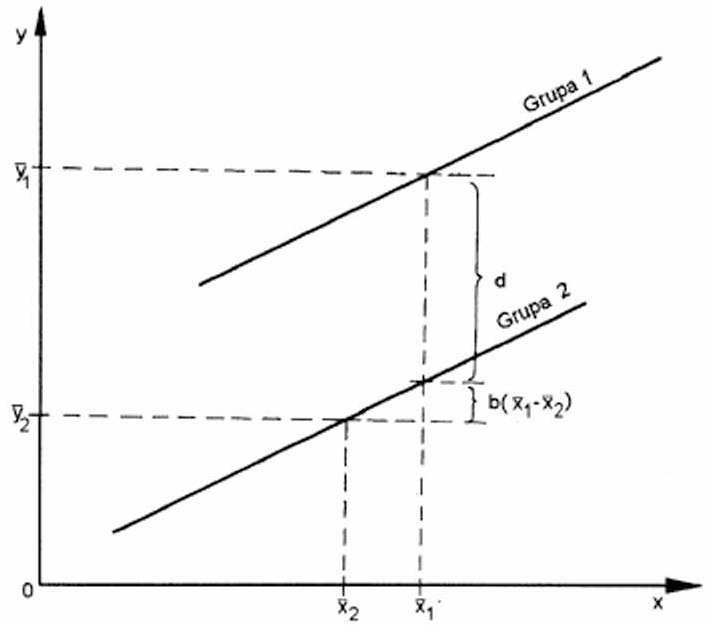
natomiast nie ma podstaw do odrzucenia hipotezy o równoległości prostych regresji

$$F_2 < {}_{0,05}F^{(2)}_{(66)} = 3{,}142$$

8.4.3 Badanie odległości pionowej dwóch prostych regresji

Następnym pytaniem, które nasuwa się po skonstatowaniu, że nie ma wyraźnych podstaw, aby uważać proste regresji w grupach za nierównoległe, to pytanie czy te równoległe proste pokrywają się czy też nie, a jeżeli nie, to jaka jest ich odległość wzajemna. Innymi słowy chodzi o to, czy różnica w położeniu prostych równoległych wynika z błędu losowego, czy też można uważać, że proste te są istotnie różne oraz jak można oszacować odległość między tymi prostymi.

Na początek rozważymy $k = 2$ grupy. Jeżeli nie ma podstaw do odrzucenia hipotezy o równoległości prostych regresji, to pionowa odległość d między tymi prostymi dana będzie wzorem (por. także rys. 8.6)



Rys. 8.6 Ilustracja sposobu obliczania odległości pionowej dwu równoległych prostych regresji.

$$d = \bar{y}_1 - \bar{y}_2 - b \cdot (\bar{x}_1 - \bar{x}_2) \tag{8.53}$$

gdzie b jest wspólnym współczynnikiem nachylenia

$$b = \frac{(Sxy)_1 + (Sxy)_2}{(Sx^2)_1 + (Sx^2)_2}$$

Wariancję wielkości d można oszacować przy pomocy zależności:

$$\sigma^2(d) = s_C^2 \cdot \left[\frac{1}{n_1} + \frac{1}{n_2} + \frac{(\bar{x}_1 - \bar{x}_2)^2}{(Sx^2)_1 + (Sx^2)_2}\right] \tag{8.54}$$

gdzie s_C^2 jest resztowym średnim kwadratem odchyleń od prostych równoległych (suma *SKMW* + *SKRW* z poprzedniego punktu podzielona przez liczbę stopni swobody)

$$s_C^2 = \frac{(Sy^2)_1 + (Sy^2)_2 - \frac{[(Sxy)_1 + (Sxy)_2]^2}{(Sx^2)_1 + (Sx^2)_2}}{n_1 + n_2 - 3} \tag{8.55}$$

Testujemy hipotezę zerową mówiącą o tym, że obie równoległe proste regresji pokrywają się, czyli ich odległość pionowa wynosi zero. Obliczamy więc statystykę t

$$t = \frac{d}{\sqrt{\sigma^2(d)}} \tag{8.56}$$

i porównujemy ją z wartością krytyczną rozkładu t-Studenta przy $n_1 + n_2 - 3$ stopniach swobody. Dalsze postępowanie jest znane. Granice przedziału ufności dla odległości pionowej oblicza się w zwykły sposób

$$d \pm \alpha\, t_{(n_1 + n_2 - 3)} \sqrt{\sigma^2(d)}$$

przy czym $1 - \alpha$ jest współczynnikiem ufności.

Przykład 8.5

Obliczymy i oszacujemy pionową odległość pomiędzy dwoma równoległymi prostymi regresji badanymi w przykładzie 8.3 (chodzi o proste uzyskane dla preparatów 1 i 2 z tabel 8.6 i 8.7). Odległość d będzie równa:

$$d = 1{,}95 - 3{,}4432 - 3{,}2791\,(1 - 0{,}8358) = -2{,}0315$$

Obliczymy teraz s_C^2 i $\sigma_2(d)$:

$$s_C^2 = \frac{85{,}175 + 59{,}4915 - \frac{(17{,}6085 + 8{,}3733)^2}{5{,}4362 + 2{,}4874}}{30 + 22 - 3}$$

$$\sigma^2(d) = 1{,}2137\left[\frac{1}{30} + \frac{1}{22} + \frac{(1 - 0{,}8358)^2}{5{,}4362 + 2{,}4874}\right] = 0{,}0098$$

Stąd

$$t = \frac{-2{,}0315}{\sqrt{0{,}0998}} = -6.4323$$

Ponieważ

$$|t| > {}_{0,05}\, t_{(49)} = 2{,}012$$

więc hipotezę o pokrywaniu się prostych regresji odrzucamy. Obliczamy 95%-owe granice przedziału ufności dla odległości pionowej dwóch prostych regresji

$$-2{,}0315 \pm 2{,}012\ \sqrt{0{,}0998} = -1{,}3961 \qquad \text{oraz} \qquad -2{,}6670$$

Znak minus przy wartości odległości d wskazuje, że prosta 2 leży „nad" prostą 1, czyli odwrotnie, niż na rysunku 8.6.

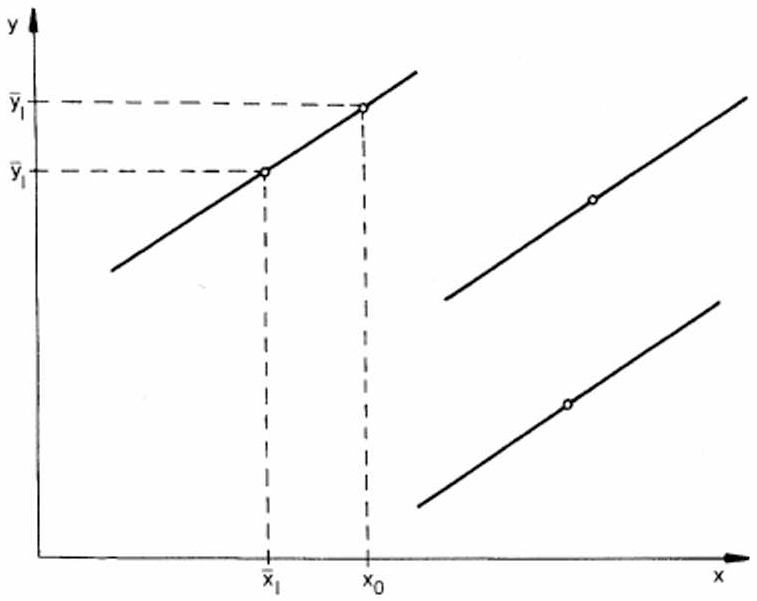
8.4.4 Badanie położenia równoległych prostych regresji dla kilku grup. Analiza kowariancji

W przypadku gdy liczba grup jest większa od 2, dla zbadania położenia wzajemnego równoległych prostych regresji stosujemy pojęcie poprawionych wartości średnich $\overline{y}'_1$. Gdyby średnia wartość zmiennej x w l-tej grupie wynosiła x_0, a nie $\overline{x}_l$, to średnia wartość zmiennej y wynosiłaby wtedy

$$\overline{y}'_l = \overline{y}_l + b \cdot (x_0 - \overline{x}_l)$$

Wartość $\overline{y}'_l$ nazywamy poprawioną wartością średnią (por. rysunek 8.7). Gdyby wszystkie równoległe proste regresji pokrywały się, wówczas wszystkie tak wyliczone (dla pewnego ustalonego x_0) poprawione wartości średnie $\overline{y}'_l$ byłyby równe.

Stawiamy hipotezę zerową, że równoległe proste regresji w grupach pokrywają się. Jest to równoważne równości poprawionych średnich i właśnie będziemy sprawdzać, czy różnice pomiędzy poprawionymi średnimi wynikają z błędu losowego, czy też poprawione średnie są istotnie różne. Różnice między kilkoma średnimi badaliśmy metodą analizy wariancji, różnice między poprawionymi średnimi testujemy używając metody zwanej analizą kowariancji. Jest to metoda zbliżona do analizy wariancji w klasyfikacji pojedynczej, wykorzystująca jednak poza odpowiednimi sumami kwadratów także sumy iloczynów typu $x \cdot y$. Sposób postępowania jest niemal identyczny jak w analizie wariancji z tym, że używa się tzw. poprawionych sum kwadratów obliczanych na podstawie odpowiednich



Rys. 8.7 Ilustracja pojęcia poprawionej średniej wartości $\bar{y}_l$' dla l-tej grupy.

wzorów, które zostaną podane dalej bez szczegółowego uzasadniania, gdyż wykraczałoby to poza ramy niniejszego skryptu.

Chcąc zbadać różnice między poprawionymi średnimi, rozbijamy ogólną poprawioną sumę kwadratów *PSK* na dwa składniki: międzygrupową poprawioną sumę kwadratów *PSKMG* i wewnątrzgrupową poprawioną sumę kwadratów *PSKWG*.

$$PSK = PSKMG + PSKWG \tag{8.57}$$

Poszczególne składniki (8.57) wyliczamy ze wzorów:

$$PSK = (Sy^2)_{całk} - \frac{\left[(Sxy)_{całk}\right]^2}{(Sx^2)_{całk}} \tag{8.58}$$

$$PSKWG = \sum_l (Sy^2)_l - \frac{\left[\sum_l (Sxy)_l\right]^2}{\sum_l (Sx^2)_l} \tag{8.59}$$

$$PSKMG = PSK - PSKWG \tag{8.60}$$

przy czym

$$(Sy^2)_{całk} = \sum_l \sum_i y_{il}^2 - \frac{\left(\sum_l \sum_i y_{il}\right)^2}{N} \tag{8.61}$$

$$(Sx^2)_{całk} = \sum_l \sum_i x_{il}^2 - \frac{\left(\sum_l \sum_i x_{il}\right)^2}{N} \tag{8.62}$$

$$(Sxy)_{całk} = \sum_l \sum_i x_{il}\, y_{il} - \frac{\left(\sum_l \sum_i x_{il}\right) \cdot \left(\sum_l \sum_i x_{il}\right)}{N} \tag{8.63}$$

$$N = \sum_l n_l \tag{8.64}$$

Na podstawie poprawionych sum kwadratów i odpowiednich liczb stopni swobody (por. tabela 8.10) obliczamy średnie poprawione kwadraty s_B^2 i s_C^2, a następnie testujemy ich stosunek F przy $k - 1$ i $N - k - 1$ stopniach swobody i poziomie istotności α. Hipotezę o pokrywaniu się równoległych prostych regresji odrzucamy, jeżeli stosunek F będzie większy lub równy wartości krytycznej.

Tabela 8.10

Tablica analizy kowariancji dla badania położenia równoległych prostych regresji w k grupach

Źródło zmienności	Poprawiona suma kwadratów	Liczba stopni swobody	Średni kwadrat	Stosunek wariancji
Pomiędzy grupami	$PSKMG$	$k - 1$	$s_B^2 = \frac{PSKMG}{r-1}$	$F = \frac{s_B^2}{s_C^2}$
Wewnątrz grup	$PSKWG$	$N - k - 1$	$s_C^2 = \frac{PSKWG}{N-k-1}$	
Całkowita	SK	$N - 2$		

Jeżeli test pozwoli na odrzucenie hipotezy o pokrywaniu się równoległych prostych regresji, to możemy obliczyć różnicę między pewnymi dwoma poprawionymi średnimi

$$\bar{y}'_p - \bar{y}'_m = \bar{y}_p - \bar{y}_m - b\,(\bar{x}_p - \bar{x}_m) \tag{8.65}$$

która będzie zarazem pionową odległością obu prostych regresji. Różnicę tę można testować obliczając statystykę t:

$$t = \frac{\bar{y}'_p - \bar{y}'_m}{\sqrt{\sigma^2\,(\bar{y}'_p - \bar{y}'_m)}} \tag{8.66}$$

gdzie

$$\sigma^2\,(\bar{y}'_p - \bar{y}'_m) = s_C^2 \cdot \left[\frac{1}{n_p} + \frac{1}{n_m} + \frac{(\bar{x}_p - \bar{x}_m)^2}{\sum\limits_l (Sx^2)_l}\right] \tag{8.67}$$

a s_C^2 jest poprawionym średnim kwadratem wewnątrzgrupowym (z tablicy analizy kowariancji). Statystyka (8.66) ma rozkład t-Studenta o $N - k - 1$ stopniach swobody. Dalsze postępowanie jest znane.

Przykład 8.6

Porównamy położenia trzech równoległych prostych regresji, będących wynikiem badań efektu terapeutycznego 3 preparatów przeciwkrzywicznych (por. przykład 8.4 oraz tabele 8.6 i 8.7). Analiza kowariancji dostarcza wyników przedstawionych w tabeli 8.11. Ponieważ

Tabela 8.11

Tablica analizy kowariancji dla analizy danych z tabeli 8.6

Źródło zmienności	Poprawiona suma kwadratów	Liczba stopni swobody	Średni kwadrat	Stosunek wariancji	Istotność
Między grupami	52,8790	2	26,4395	19,0117	$P < 0{,}005$
Wewnątrz grup	94,5671	68	1,3907		
Całkowita	147,4461	70			

$$F = 19{,}0117 > {}_{0{,}005}F^{(2)}_{(68)}$$

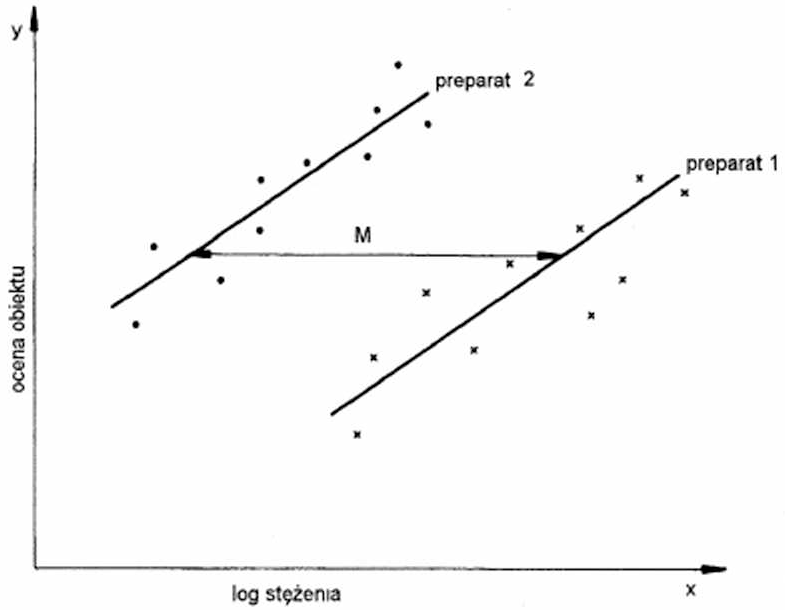
więc hipotezę o pokrywaniu się prostych regresji można odrzucić twierdząc, że istotność różnic w położeniu prostych jest wysoka.

8.4.5 Badanie poziomej odległości równoległych prostych regresji dla potrzeb badań biologicznych

Często w badaniach biologicznych interesuje nas obliczenie i wytestowanie poziomej odległości między równoległymi prostymi regresji. Jeżeli dysponujemy pomiarami efektywności terapeutycznej dwóch preparatów w zależności od logarytmu dawki, a dokładniej logarytmu stężenia (gdzie stężenie jest dawką preparatu na jednostkę wagi ciała) i stwierdziliśmy poprzez odpowiednie testy statystyczne, że:

— zależność może być uznana za prostoliniową,
— linie regresji dla obu preparatów można uważać za równoległe,
— proste regresji nie pokrywają się,

to szczególnie będziemy chcieli wyznaczyć poziomą odległość między prostymi, gdyż ma ona oczywistą interpretację. Odległość pozioma M (patrz rys. 8.8) jest różnicą logarytmów stężeń preparatu koniecznych do uzyskania tego samego efektu terapeutycznego, czyli (po zastosowaniu przekształcenia odwrotnego) ilorazem takich stężeń obu preparatów, przy których uzyskujemy ten sam efekt leczniczy. Na ogół w badaniach tego typu jeden z preparatów traktujemy jako standardowy (kontrolny). Takim preparatem może być np. preparat 1 z rys. 8.8. W badaniu zależy nam na oszacowaniu liczby R takiej, że



Rys. 8.8 Ilustracja poziomej odległości między dwoma równoległymi prostymi regresji.

$$M = \log R$$

którą nazywamy stosunkiem mocy preparatu testowanego (np. preparat 2 z rys. 8.8) w porównaniu do preparatu standardowego. Różnica pozioma M jest bardziej predystynowana do wykorzystania w badaniach niż pionowa, gdyż w przypadkach ewentualnych nieliniowości rzeczywistej zależności efektu od logarytmu stężenia, jeżeli charakter nieliniowości jest jednolity dla obu krzywych regresji (w sensie przesunięcia poziomego), to pozioma odległość na skali logarytmu stężenia będzie dalej stała, mimo że odległość pionowa może nie być zupełnie niezmienna. Aby uzyskać prostoliniowość regresji w przypadku znacznych nieliniowości, należy stosować odpowiednie przekształcenia danych na osi rzędnych (por. rozdział 2).

Pozioma odległość M może być wyliczona jako

$$M = \bar{x}_1 - \bar{x}_2 - \frac{\bar{y}_1 - \bar{y}_2}{b} \tag{8.68}$$

gdzie b jest wspólnym współczynnikiem nachylenia prostych równoległych. Przybliżony wzór na wariancję M ma postać:

$$\sigma^2(M) = \frac{s_c^2}{b^2}\left[\frac{1}{n_1} + \frac{1}{n_2} + \frac{(M - \bar{x}_1 - \bar{x}_2)^2}{(Sx^2)_1 + (Sx^2)_2}\right] \tag{8.69}$$

gdzie s_C^2 jest średnim kwadratem resztowym wyrażonym wzorem (8.55). Przed wykonaniem testów statystycznych obliczamy wartość g

$$g = \frac{\alpha\, t^2_{(n_1+n_2-3)} \cdot s_C^2}{b^2 \cdot \left[(Sx^2)_1 + (Sx^2)_2\right]} \tag{8.70}$$

Jeżeli $g < 0{,}1$ to można testować wartość M wykorzystując statystykę t

$$t = \frac{M}{\sqrt{\sigma^2(M)}} \tag{8.71}$$

i porównując jej moduł z wartością krytyczną $\alpha\, t_{(n_1+n_2-3)}$

Wtedy granice przedziału ufności dla odległości pionowej M będą określone jako

$$M \pm \alpha\, t_{(n_1+n_2-3)} \cdot \sqrt{\sigma^2(M)} \tag{8.72}$$

Gdy $0,1 < g < 1$, wówczas dla wyznaczenia granic przedziału ufności należy skorzystać z poniższej formuły:

$$\frac{M \pm \frac{\alpha t_{(n_1+n_2-3)} s_C}{b} \sqrt{(1-g)\cdot\left(\frac{1}{n_1}+\frac{1}{n_2}\right)+\frac{(M-\bar{x}_1-\bar{x}_2)_2}{(Sx^2)_1+(Sx^2)_2}}}{1-g}$$

Gdy zaś $g > 1$ należy się spodziewać, że wspólne nachylenie b prostych regresji nie jest istotnie różne od zera i dalsza analiza jest bezprzedmiotowa.

Przykład 8.7.

Zbadamy poziomą odległość dwóch równoległych prostych regresji rozważanych w przykładach 8.3 i 8.5 (por. także tabele 8.6 i 8.7). Obliczono wtedy

$$s_C^2 = 1,2137$$

$$b = 3,2791$$

Przyjmujemy poziom istotności $\alpha = 0,05$, odczytujemy wartość krytyczną rozkładu t

$$_{0,05}t_{(49)} = 2,012$$

i wyznaczamy wartość g

$$g = \frac{(2,012)^2 \cdot 1,2137}{(3,2791)^2 \cdot (5,4361 + 2,4874)} = 0,0577$$

Obliczamy także odległość poziomą M

$$M = 1,0 - 0,8358 - \frac{1,95 - 3,4432}{3,2791} = 0,6195$$

Ponieważ $g < 0,1$ możemy obliczyć $\sigma^2(M)$ ze wzoru (8.69) i wykorzystać później (8.71) i (8.72). Mamy więc

$$\sigma^2(M) = \frac{1,2137}{(3,2791)^2}\left[\frac{1}{30}+\frac{1}{22}+\frac{(0,6195-1,0+0,8358)^2}{5,4361+2,4874}\right] = 0,01185$$

$$t = \frac{0,6195}{\sqrt{0,01185}} = 5,692$$

Ponieważ

$$|t| > {}_{0,05}t_{(49)}$$

więc wnioskujemy, że odległość pozioma jest istotnie różna od zera. 95%-owe granice przedziału ufności są następujące:

$$0{,}6191 \pm 2{,}102 \cdot \sqrt{0{,}01185} = 0{,}4006 \quad \text{oraz} \quad 0{,}8385$$

Ponieważ

$$0{,}4006 = \log 2{,}515$$

$$0{,}6195 = \log 4{,}164$$

$$0{,}8385 = \log 6{,}895$$

więc stosunek mocy R badanych preparatów może być oszacowany jako

$$2{,}515 < R < 6{,}895$$

9. TESTY NIEPARAMETRYCZNE

9.1 Efektywność testów

Istnieje rozpowszechnione przekonanie, że testy nieparametryczne są „mniej skuteczne" od testów parametrycznych. Jest to przeświadczenie w gruncie rzeczy prawidłowe, gdyż wymagając słabszych założeń odnośnie charakteru danych (na przykład akceptując dane jakościowe na równi z ilościowymi) oraz nie wprowadzając ograniczeń co do charakteru rozkładu, jakiemu podlegają te dane — testy rozważanego tu typu muszą dostarczać mniej precyzyjnych odpowiedzi i rozstrzygnięć. Jednak odpowiednio dobrany i umiejętnie zastosowany test nieparametryczny może dostarczyć równie wartościowych podstaw do wnioskowania statystycznego, jak klasyczne testy *chi*-kwadrat czy *t*-Studenta. Na ogół jednak dla osiągnięcia takiej samej mocy testu konieczne jest użycie większej liczby obserwacji w teście nieparametrycznym, niż ta, która wystarcza w teście parametrycznym. W związku z tym mówi się zwykle o efektywności testu nieparametrycznego, wyrażając ją procentowym stosunkiem wielkości próby dla testu parametrycznego o tej samej mocy.

Przykład.

Stwierdzono, że porównanie dwóch sposobów leczenia *gruczolaka stercza*, wykonane testem *t*-Studenta ujawniło statystycznie znamienną różnicę na korzyść zabiegu operacyjnego przy założonym poziomie istotności $\alpha = 0{,}05$. Obalenie hipotezy zerowej nastąpiło w oparciu o porównanie średnich z dwóch równolicznych zbiorów obserwacji klinicznych o liczebności $N_t = 100$ pacjentów w każdej grupie. Prawdopodobieństwo błędu II rodzaju (to znaczy błędu nie odrzucenia H_0 mimo jej fałszywości) oszacowane zostało na poziomie $\beta = 0{,}12$, zatem moc testu Studenta w rozważanym przykładzie wyniosła $(1 - \beta) = 88\%$. Następnie te same wyniki poddano ocenie za pomocą testu nieparametrycznego, ponieważ prawdziwość założenia, że porównywane dane pochodzą z populacji o rozkładzie normalnym wydawała się problematyczna. Zastosowano test *U*-Manna-Whitneya (opisany dalej) i ponownie uzyskano efekt w postaci zdecydowanego odrzucenia H_0 na poziomie $\alpha = 0{,}05$, ale tym razem oszacowanie błędu drugiego rodzaju dało wartość $\beta = 0{,}21$, czyli moc testu nieparametrycznego wynosiła zaledwie $(1 - \beta) = 79\%$. Ponieważ rozporządzano dostatecznie licznym zbiorem historii chorób

dużego ośrodka onkologicznego, zaczęto poszerzać próbę, wprowadzając dalsze dane i kontrolując, jak wzrost N wpływa na wartość β. Okazało się, że porównywalną z testem Studenta moc testu Manna-Whitneya uzyskano przy N_u = 150 (oszacowanie β wynosiło wtedy 0,11 czyli moc osiągała poziom 89%).

Z definicji efektywności można więc stwierdzić, że efektywność testu U w stosunku do testu t oszacować można na

$$E = \frac{N_t}{N_u} = \frac{100}{150} = 75\%$$

Warto dodać, że wyliczona w przykładzie wartość efektywności testu Manna-Whitneya jest mniejsza od podawanej w literaturze wartości e = 95%.

9.2 Porównywanie populacji

9.2.1 Ogólna charakterystyka zadania

Omawiana w tym podrozdziale grupa testów nieparametrycznych, służących do porównywania populacji, pełni w zasadzie podobną rolę, jak rozważany poprzednio test t-Studenta. Poszukujemy odpowiedzi na pytanie: czy dwie rozważane populacje różnią się w sposób istotny, czy nie. O tym, że się różnią — upewnia nas wstępna analiza wyników (jeśli takiej różnicy nie dostrzegamy, wówczas bezcelowa jest cała analiza matematyczna, bo nie ma czego badać). O tym, że różnicy takiej oczekujemy, decyduje sposób rekrutacji danych do obydwu rozważanych populacji — najczęściej jedna z nich obejmuje pacjentów stanowiących przedmiot obserwacji, zaś druga jest grupą kontrolną, stanowiącą punkt odniesienia. Pytanie jest zatem tylko jedno: czy obserwowana różnica jest przejawem jakiejś głębszej prawidłowości, czy też wynikła z przypadkowego zbiegu okoliczności?

W teście t-Studenta zakładaliśmy, że charakter rozkładu jest określony (normalny) oraz że badane rozkłady są zgodne we wszystkim z wyjątkiem różnicy wartości oczekiwanych. Obecnie żadnych takich założeń nie robimy, stawiając pytanie maksymalnie szeroko: czy obserwowane rozkłady różnią się, czy nie? Poznamy niżej kilka testów, przy pomocy których można poszukiwać odpowiedzi na to pytanie.

9.2.2 Test Walda-Wolfowitza (test serii)

Zaletą testu Walda-Wolfowitza jest prostota jego przeprowadzania, natomiast wadą — stosunkowo niska efektywność. W aktualnej sytuacji, kiedy strona rachunkowa przeprowadzanych rozważań obciąża z reguły jedynie komputery a zawiłości algorytmu wbudowane są w używany program — test ten ma stosunkowo mniejsze znaczenie i omawiamy go

tu głównie dlatego, że jest on w wielu klasycznych i uznawanych do dziś pozycjach literatury rekomendowany jako podstawowy.

Zakładamy, że dane podlegające ocenie w rozważanym teście mogą być między sobą porównywane w kategoriach „mniejsze — większe" i że ich wartości można w związku z tym uporządkować, na przykład od najmniejszej do największej. Warto zwrócić uwagę, że jest to na ogół znacznie słabsze założenie, niż wymagane w teście *t*-Studenta wyrażenie danych w postaci konkretnych wartości liczbowych. Niech będzie danych N_1 obserwacji pochodzących z pierwszej wyróżnionej populacji oraz N_2 obserwacji z drugiej populacji. Uszeregowawszy te obserwacje według porządku narzuconego przez rozważane dane możemy stwierdzić, że występować będą **serie** danych pochodzących z obu rozważanych grup. Istota testu polega na policzeniu, ile tych serii występuje. Jeśli jest ich mało — to mamy prawo przypuszczać, że czynnik powodujący wyróżnienie dwóch grup: danych badanych i grupy kontrolnej okazał znaczny wpływ na wartość ocenianego parametru. Jeśli natomiast serii tych jest dużo — wówczas niemożliwe jest odrzucenie hipotezy H_0 głoszącej, że rozważane populacje nie różnią się od siebie. Wszak dokładne „wymieszanie" wartości analizowanego parametru (czego znamiennym dowodem jest duża liczba wykrytych serii) stanowi argument na korzyść hipotezy H_0!

Tak więc w teście Walda-Wolfowitza trzeba tylko wyznaczyć najprostszą statystykę w postaci zliczonej liczby serii pomiarowych (oznaczanej ζ) i już można odwołać się do tablicy wartości krytycznych, w której zebrano (dla założonego poziomu istotności α oraz dla różnych wartości liczebności N_1 i N_2) graniczne wartości ${}_\alpha\zeta(N_1, N_2)$. Jeśli wyliczone z obserwacji $\zeta < {}_\alpha\zeta(N_1, N_2)$ wówczas jest podstawa do odrzucenia H_0, w przeciwnym przypadku należy wstrzymać się od osądu (Jak pamiętamy, zawsze — a szczególnie w przypadku testów nieparametrycznych — brak podstaw do odrzucenia hipotezy zerowej H_0 **nie upoważnia** do jej **przyjęcia!**).

9.2.3 Przykład użycia testu Walda-Wolfowitza

Rozważmy konkretny przykład. Oceniano stan kliniczny 19 pacjentów cierpiących na przewlekłe zapalenie górnych dróg oddechowych. Lekarze określili ten stan, wskazując pozycję każdego pacjenta na liście od pierwszego (najlżejszy) do ostatniego (najcięższy stan) miejsca w tym swoistym szeregu rang. Niezależnie od tego mikrobiologowie ustalili dla każdego pacjenta rodzaj drobnoustrojów chorobotwórczych, dominujących w wymazie z nosogardzieli, dzieląc w ten sposób populację pacjentów na dwie podgrupy: u jednych stwierdzono obecność gronkowców ($N_1 = 10$) a u drugich paciorkowców ($N_2 = 9$). Należy zbadać, czy hipoteza, że paciorkowce są odpowiedzialne za cięższe kliniczne przypadki znajdujące potwierdzenie w obserwacjach. Oczywiście na początek stawiamy hipotezę H_0 , że obie populacje nie różnią się i analizujemy przytoczoną niżej tabelę 9.1 aby ustalić, czy można ją obalić. Jak wynika z tabeli $\zeta = 12$, natomiast znaleziona w tabeli (na końcu skryptu) wartość krytyczna ${}_\alpha\zeta(N_1, N_2)$ dla $\alpha = 0{,}05$,

$N_1 = 10$ i $N_2 = 9$ wynosi 6. Jak z tego wynika $\zeta > {}_{\alpha}\zeta(N_1, N_2)$ z czego wynika, że brak podstaw do obalenia hipotezy H_0.

Tabela 9.1

Zestawienie oceny stanu klinicznego pacjentów z przewlekłym zapaleniem górnych dróg oddechowych wraz z informacją, jaki rodzaj bakterii wykryto w wymazie z nosogardzieli

Ocena stanu klinicznego (ranga)	Rodzaj bakterii g = gronkowce p = paciorkowce	Numer serii
1	g	1
2	g	
3	p	2
4	g	3
5	g	
6	g	
7	g	
8	p	4
9	g	5
10	p	6
11	g	7
12	p	8
13	p	
14	g	9
15	p	10
16	p	
17	g	11
18	p	12
19	p	

Używanie tabeli dla wartości krytycznych ${}_{\alpha}\zeta(N_1, N_2)$ jest uciążliwe dla większych wartości N_1 i N_2. Na szczęście rozkład ζ zmierza szybko (dla dużych N_1 i N_2) do rozkładu normalnego. Parametrami tego rozkładu są:

$$\mu_z = \frac{2N_1 N_2}{N_1 + N_2} + 1 \qquad \sigma_z^2 = \frac{2N_1 N_2 (2N_1 N_2 - N_1 - N_2)}{(N_1 + N_2)^2 (N_1 + N_2 - 1)}$$

Posługując się zmienną

$$\overline{\zeta} = \frac{\zeta - \mu_z}{\sigma_z}$$

możemy więc zakładać, że jest to zmienna mająca standaryzowany rozkład normalny i możemy ją oceniać w znany z praktyki testów parametrycznych, typowy sposób.

Wprawdzie w rozważanym wyżej przykładzie wartościi N_1 i N_2 trudno było uznać za „duże", jednak dla ilustracji można sprawdzić, do jakich wniosków doprowadzi w tym wypadku założenie o normalności rozkładu ζ. Zaczniemy od wyliczenia parametrów rozkładu

$$\mu_z = \frac{2 * 10 * 9}{10 + 9} = 10{,}47$$

$$\sigma_z = \sqrt{\frac{2 * 10 * 9 * (2 * 10 * 9 - 10 - 9)}{19^2 * 18}} = 2{,}11$$

$$\overline{\zeta} = (12 - 10{,}47) / 2{,}11 = 0{,}725$$

Wynik oceny możliwości obalenia H_0 jest w tym wypadku natychmiastowo negatywny, bez konieczności odwoływania się do jakichkolwiek tablic: otóż $\zeta > \mu_z$ ($\overline{\zeta} > 0$), a tymczasem warunkiem odrzucenia H_0 jest $\overline{\zeta} < 0$. Jak z tego wynika, wniosek z obliczeń opartych na przybliżeniu rozkładem normalnym potwierdza w rozważanym przykładzie wniosek z obliczeń dokładnych.

9.2.4 Test Manna–Whitneya

Omówiony wyżej test Walda-Wolfowitza miał istotne zalety w postaci prostego i łatwego do realizacji algorytmu, miał jednak także wadę w postaci małej efektywności. Niektórzy badacze [Blalock] szacują ją nawet zaledwie na 75%, co jest wynikiem znacznie gorszym niż rezultaty uzyskiwane w innych testach. Pod tym względem znacznie korzystniejszy jest test Manna-Whitneya, któremu przypisuje się efektywność 95%. Test ten opisywany jest w podręcznikach w różnej postaci i przy prezentacji istoty tego testu

w niniejszym skrypcie wybrano arbitralnie jedną z tych alternatywnych form opisu, kierując się głównie łatwością komputerowej implementacji testu. W uzupełnieniu tych wprowadzających uwag warto tylko jeszcze odnotować fakt, że test Manna-Whitneya w niektórych swoich odmianach nazywany jest w literaturze testem Wilcoxona — chociaż bez trudu można wykazać merytoryczną równoważność wszystkich tych formalizacji.

Punktem wyjścia w opisywanym teście jest znowu zbiór danych — ilościowych lub jakościowych — pochodzących z dwóch porównywanych populacji. Wśród danych tych musi być wprowadzone uporządkowanie od najmniejszej do największej wartości — podobnie jak w omawianym wyżej teście serii (Walda-Wolfowitza). Jednak możliwe jest tu wprowadzenie tzw. rang wiązanych, to znaczy dopuszczalna jest sytuacja, w której kilka obserwacji otrzyma tę samą ocenę i zostanie sklasyfikowane na „tym samym miejscu" w tabeli rang. Ułatwia to na ogół rolę eksperta, który musi dokonać uporządkowania obserwacji względnie umożliwia uwzględnienie sprzecznych opinii kilku rzeczoznawców. Stanowi to — obok wyższej efektywności — kolejny atut testu Manna-Whitneya i zachęca do jego szerokiego stosowania.

Podstawą do podjęcia decyzji jest w opisywanym teście **suma** rang (ocen, pozycji na liście) uzyskanych przez obydwie porównywane populacje. Ponieważ łączna ilość rang pozostających do rozdysponowania jest ograniczona, gdyż przyjmują one wartości z przedziału od 1 do N, gdzie $N = N_1 + N_2$ jest łączną liczbą wszystkich obserwacji, przeto wystarczy policzyć sumę rang dla jednej tylko populacji, gdyż drugą sumę można wyliczyć ze wzoru

$$R_2 = \frac{N(N+1)}{2} - R_1$$

gdzie R_1 i R_2 są odpowiednimi sumami rang dla obydwu rozważanych populacji.

Przykład

Sposób obliczania R_1 i R_2 łatwo można prześledzić w oparciu o dane zebrane w tabeli 9.1. Policzmy sumę rang dla populacji w której wykryto gronkowca

$$R_1 = 1 + 2 + 4 + 5 + 6 + 7 + 9 + 11 + 14 + 17 = 76$$

Policzmy także sumę rang dla drugiej grupy, tzn. dla paciorkowców:

$$R_2 = 3 + 8 + 10 + 12 + 13 + 15 + 16 + 18 + 19 = 114$$

a następnie skorzystajmy z przytoczonego wyżej wzoru:

$$R_2 = 19 * 20/2 - 76 = 190 - 76 = 114$$

Rzeczywiście, zgadza się.

Mając zliczone wartości R_1 i R_2, a także znając liczebność populacji N_1 i N_2 (jak pamiętamy w rozważanym przykładzie $N_1 = 10$ a $N_2 = 9$) możemy wyliczyć statystykę testu Manna-Whitneya, oznaczaną zwyczajowo U. Mamy do dyspozycji dwa równoważne wzory

$$U = N_1 * N_2 + \frac{N_1 (N_1 + 1)}{2} - R_1$$

$$U = N_1 * N_2 + \frac{N_2 (N_2 + 1)}{2} - R_2$$

Przykład

Korzystając z drugiego z podanych wzorów wyliczamy wartość statystyki U dla rozważanego wyżej przykładu

$$U = 10 * 9 + 9 * 10/2 - 114 = 21$$

Teraz zachodzi potrzeba porównania wyliczonej wartości U z wartością krytyczną. Wartość ${}_{\alpha}U(N_1, N_2)$ można odczytać z odpowiednich tabel, podanych na końcu skryptu, przy czym dla $\alpha = 0{,}05$ i warunków określonych w rozważanym przykładzie wartość krytyczna wynosi ${}_{\alpha}U = 20$. Hipotezę H_0 można obalić, jeśli obliczona na podstawie danych eksperymentalnych wartość statystyki U jest **mniejsza** od krytycznej, zatem w omawianym wypadku wniosek z zastosowania testu Manna-Whitneya jest analogiczny jak uprzednio określony przy zastosowaniu testu Walda-Wolfowitza: brak podstaw do obalenia hipotezy H_0, zatem brak również podstaw do przypuszczeń, że chorzy zakażeni przez paciorkowce ciężej przechodzą zapalenie górnych dróg oddechowych. Zwróćmy jednak uwagę jak mało brakowało do tego, by decyzja była przeciwna. Ta subtelność testu Manna-Whitneya potwierdza jego renomę jako testu o wysokiej efektywności. Istotnie, prawdopodobieństwo błędu II rodzaju β jest tu wyraźnie niższe niż w przypadku testu serii, gdyż po prostu hipotezę H_0 znacznie łatwiej obalić.

Mankamentem testu Manna-Whitneya jest — podobnie jak w przypadku testu serii — konieczność stosowania skomplikowanych tablic wartości krytycznych. Dlatego znaczną popularność zdobyła sobie uproszczona wersja tego testu, nazywana testem Z. Zasadniczo test Z powinno się stosować dla dużych wartości N_1 i N_2 (>20), jednak jego użycie dla mniejszych N daje często także dobre rezultaty. Test Z polega na wyliczeniu statystyki Z określonej wzorem

$$Z = \frac{R_1 - R_2 - (N_1 - N_2)(N + 1)/2}{\sqrt{N_1 N_2 (N + 1)/3}}$$

Statystyka Z ma (w przybliżeniu) standaryzowany rozkład normalny, można więc oceniać jej wartości poprzez porównanie z wartościami odczytanymi z tabeli dla rozkładu normalnego (przykładowo dla typowo przyjmowanego $\alpha = 0{,}05$ wartość krytyczna wynosi oczywiście ${}_{\alpha}Z = 1{,}96$). Zasada wnioskowania statystycznego obowiązująca przy teście Z polega na odrzuceniu H_0, jeżeli wartość Z jest **większa** (biorąc pod uwagę bezwzględną wartość) od wartości krytycznej ${}_{\alpha}Z$. Przy okazji warto zauważyć, że struktura wzoru dla Z ujawnia, o co właściwie we wszystkich tych testach chodzi. Zauważmy, że podstawowym elementem wzoru dla Z jest różnica $R_1 - R_2$, co żywo przypomina licznik wzoru dla testu t-Studenta porównującego dwie średnie.

Przykład

Skorzystajmy z testu Z i raz jeszcze poddajmy ocenie dane zawarte w tabeli 9.1. Ponieważ wszystkie elementy wzoru są znane, wystarczy porachować wartość statystyki Z:

$$Z = \frac{76 - 114 - (10 - 9) * 20/2}{\sqrt{10 * 9 * 20/3}} = -1{,}96$$

Jak widać wartość statystyki wypadła dokładnie na granicy znamienności, co jednak nie upoważnia jeszcze do obalenia hipotezy H_0. Nie zapominajmy przy tym, że jest to postępowanie przybliżone, dające dobre oszacowanie jedynie dla dużych wartości liczebności N_1 i N_2, co w rozważanym tu zadaniu nie zachodzi. Przedwczesne więc byłoby zamykanie tych rozważań konkluzją, że test Z jest jeszcze bardziej efektywny od „pełnego" testu U Manna-Whitneya.

Warto natomiast wskazać, że dla licznych danych, dla których zarezerwowane jest stosowanie testu Z, metodyka obliczania R_1 i R_2 staje się nieco uciążliwa, przy czym nie chodzi tu o uciążliwości obliczeniowej natury (gdyż tymi wobec powszechnej dostępności komputerów możemy się nie przejmować), lecz o trudności z samym „ustawieniem danych". Jeśli bowiem można myśleć o uporządkowaniu w jednolitą tabelę oceny stanu klinicznego kilkunastu pacjentów, to trudno oczekiwać, że ktoś dokona takiej klasyfikacji dla kilkuset osobników! Dlatego przy niezaprzeczalnej użyteczności testu Manna-Whitneya dla nielicznych obserwacji (bardzo typowych przy opracowywaniu nowych metod diagnozowania lub doskonalszych technik leczenia) — trudno zgodzić się z wysoką oceną jego przydatności dla zadań wnioskowania statystycznego w epidemiologii, gdzie wchodzą w grę setki obserwacji. Potrzebne są więc dalsze testy.

9.2.5 Test Kołmogorowa–Smirnowa

Podstawą testu Kołmogorowa-Smirnowa jest założenie, że stosunkowo liczne dane obserwacyjne, pochodzące z dwóch populacji, wyrażają się jakościowymi parametrami, którym można przyporządkować kilka kategorii. Korzystne jest,jeśli kategorie te wpro-

wadzają pewien porządek (w rozumieniu „lepiej — gorzej" lub „więcej — mniej") ale nie jest to w rozważanej tu metodzie tak ważne, jak w dwóch uprzednio dyskutowanych testach. Dane obserwacyjne grupuje się początkowo w tabeli, której wiersze odpowiadają wyróżnionym kategoriom ocenianego parametru, a kolumny — dwu wydzielonym populacjom. Struktura takiej tabeli w ogólnym przypadku jest następująca:

$$\begin{array}{cc} x_{11} & x_{21} \\ x_{12} & x_{22} \\ x_{13} & x_{23} \\ \cdot & \cdot \\ \cdot & \cdot \\ x_{1n} & x_{2n} \end{array}$$

przy czym oczywiście

$$\sum_{i=1}^{n} x_{1i} = N_1 \qquad \text{oraz} \qquad \sum_{i=1}^{n} x_{2i} = N_2$$

gdzie N_1 i N_2 są (dużymi z reguły) liczebnościami danych z pierwszej i drugiej populacji.

Pierwszym krokiem jest dokonanie kumulacji danych w wyniku czego powstaje tabela

$$\begin{array}{cc} y_{11} & y_{21} \\ y_{12} & y_{22} \\ y_{13} & y_{23} \\ \cdot & \cdot \\ \cdot & \cdot \\ y_{1n} & y_{2n} \end{array}$$

przy czym

$$y_{ij} = \sum_{\mu=1}^{j} x_{i\mu} \qquad (i = 1, 2)$$

Oczywiście w wyniku takiego przekształcenia $y_{i1} = x_{i1}$ oraz $y_{in} = N_i$ (dla i = 1, 2).

Kolejnym przekształceniem jest zamiana skumulowanych zliczeń y_{ij} na częstości p_{ij} zgodnie ze wzorem:

$$p_{ij} = \frac{y_{ij}}{N_i}$$

Podstawą wnioskowania statystycznego w teście Kołmogorowa-Smirnowa jest statystyka D określona jako

$$D = \max_{1 \le j \le n} \left| p_{1j} - p_{2j} \right|$$

Prowadząc wnioskowanie z użyciem tego testu, stawia się (jak zwykle) hipotezę H_0 głoszącą, że pomiędzy rozważanymi populacjami nie ma żadnej różnicy. Hipotezę tę wolno obalić, jeśli wyznaczona wartość D będzie większa niż wartość krytyczna ${}_{\alpha}D$, obliczana ze wzoru

$${}_{\alpha}D = \sqrt{\frac{N_1 + N_2}{N_1 * N_2}}$$

Dla $\alpha = 0{,}05$ obowiązuje ${}_{\alpha}D = 1{,}36$, dla $\alpha = 0{,}01$ ${}_{\alpha}D = 1{,}63$, a dla $\alpha = 0{,}001$ ${}_{\alpha}D = 1{,}95$, zatem wzrost wymaganej dokładności wnioskowania powoduje wzrost ${}_{\alpha}D$, zaś zmniejszenie wymaganego poziomu ufności pociąga za sobą zmniejszenie wartości ${}_{\alpha}D$ — na przykład dla $\alpha = 0{,}10$ mamy ${}_{\alpha}D = 1{,}22$.

Metodyka stosowania testu Kołmogorowa-Smirnowa jest wyjątkowo prosta, jak łatwo zauważyć nie wymaga nawet korzystania z tablic wartości krytycznych. Potwierdza to także przedstawiony niżej przykład.

9.2.6 Przykład wykorzystania testu Kołmogorowa–Smirnowa

Przedmiotem badania jest wpływ antybiotykoterapii na wynik leczenia pewnej klasy schorzeń wirusowych. Jak wiadomo antybiotyki nie niszczą wirusów, przeto ich działanie podczas terapii chorób wirusowych musi być rozpatrywane wyłącznie jako funkcja osłonowa — aby uniemożliwić rozwinięcie się skojarzonego z infekcją wirusową zakażenia bakteryjnego. Jednak antybiotyki nie są obojętne dla zdrowia i wielu lekarzy kwestionuje celowość ich stosowania w zakażeniach wirusowych upatrując w nich czynnik pogarszający stan pacjenta. Aby rozstrzygnąć ten problem przeprowadzono obserwacje na grupie 236 chorych, którym nie podano antybiotyków i 274 chorych, u których stosowano antybiotykoterapię. Stwierdzono w każdej grupie pewną liczbę polepszeń i pewną liczbę pogorszeń stanu obserwowanych pacjentów. Z danych tych zdawało się wynikać, że antybiotykoterapia jest celowa: w grupie chorych którym podano antybiotyki obserwowano mniejszą liczbę pacjentów z pogorszeniem stanu i większą liczbę całkowitych ozdrowieńców niż w porównywanej grupie bez antybiotykoterapii. Zachodzi jednak pytanie, czy efekt ten można uznać za statystycznie znamienny? Dla uzyskania odpowiedzi na to pytanie sporządzono systematyczną tabelę obserwacji (Tabela 9.2).

Tabela 9.2

Obserwacje wyników leczenia chorych z wirusowym zapaleniem płuc w zależności od podawania lub niepodawania antybiotyków (w tabeli zebrano liczby odpowiednich pacjentów)

Lp.	efekt leczenia	bez antybiotyku	z antybiotykiem
1	pogorszenie	58	31
2	stan bez zmian	51	46
3	niewielkie polepszenie	47	53
4	wyraźne polepszenie	44	73
5	zanik objawów choroby	22	51
6	całkowite wyleczenie	14	20
	Razem	236	274

Na podstawie tej tabeli dokonano obliczeń zgodnie z zasadami obowiązującymi dla testu Kołmogorowa-Smirnowa. Wyniki tych obliczeń zebrano w tabeli 9.3, z której wynika, że wartość statystyki D wynosi w rozpatrywanym przypadku 0,187.

Tabela 9.3

Etapy przkształcania danych z tabeli 9.2 zgodnie z algorytmem testu Kołogorowa-Smirnowa

Lp.	bez antybiotyku		z antybiotykiem		Różnica	Uwagi
	skumulowane	częstość	skumulowane	częstość		
1	59	0,246	31	0,113	0,133	
2	109	0,462	77	0,281	0,181	
3	156	0,661	130	0,474	0,187	max
4	200	0,847	203	0,741	0.106	
5	222	0,941	254	0,927	0,014	
6	236	1,000	274	1,000	—	

Znajdujemy wartość krytyczną dla $\alpha = 0{,}01$:

$$_{\alpha}D = \sqrt{\frac{236 + 274}{236 * 274}} = 1{,}63 \cdot 0{,}0888 = 0{,}145$$

Okazuje się, że $D > {}_{\alpha}D$ co upoważnia nas do podjęcia decyzji o odrzuceniu hipotezy H_0 głoszącej, że antybiotykoterapia nie ma wpływu na efekt leczenia. Udowodniono w ten sposób, że z bardzo wysokim poziomem ufności ($1 - \alpha = 0{,}99$) można wykazać pozytywny wpływ osłonowego podawania antybiotyków w przypadku rozważanej klasy zakażeń wirusowych.

9.2.7 Test Wilcoxona dla par

Podobnie jak w przypadku testu Studenta, do którego stale nawiązujemy, w testach nieparametrycznych także można osiągnąć znaczne zwiększenie czułości i selektywności testu jeśli wykorzysta się dodatkowe informacje o porównywanych danych. Najbardziej typowym przykładem takiej dodatkowej informacji jest stwierdzenie, że porównywane dane tworzą naturalne pary i porównania należy dokonywać w obrębie tych par. W zakresie testów nieparametrycznych wykorzystywanym testem jest w tym wypadku test Wilcoxona. Wymaga on zastosowania danych określanych jako przynajmniej **porządkowo-metryczne**, czyli pośrednich pomiędzy danymi w pełni ilościowymi, a danymi całkowicie jakościowymi. Chodzi mianowicie o to, by nie tylko dane można było uporządkować według kategorii „mniejsze — większe", ale w dodatku aby w podobny sposób można było uporządkować **różnice** między danymi w obrębie rozważanych par.

Istota testu polega na następującej metodzie postępowania. Gromadzimy obserwacje mające postać par (x_i, y_i), $i = 1, ..., N$ — najczęściej dwa różne oznaczenia tego samego parametru u tego samego pacjenta — na przykład przed i po zabiegu operacyjnym. Następnie dla każdej pary wyznaczamy różnicę

$$d_i = x_i - y_i, \qquad i = 1, \ldots, N$$

Warto podkreślić, że zarówno x_i, y_i jak i d_i nie muszą być danymi liczbowymi w ścisłym tego słowa znaczeniu. Wystarczy jeśli można będzie sensownie zdefiniować różnicę d_i oraz dokonać pomiędzy tymi różnicami porównania, wprowadzając ich uporządkowanie i odpowiednie numery (rangi). Rangę różnicy d_i oznaczymy przez r_i przy czym podczas rangowania nie dokonuje się różnicowania pomiędzy dodatnimi i ujemnymi różnicami d_i. Natomiast wyliczając statystykę omawianego testu, oznaczoną T, musimy zesumować rangi różnic osobno dla różnic dodatnich i osobno dla różnic ujemnych. Powstają w ten sposób dwie liczby — z których do dalszych rozważań wybieramy mniejszą.

$$T = \min\left(\sum_{d_i>0} r_i, \sum_{d_i<0} r_i\right)$$

To właśnie mniejsza z sum rang jest wartością statystyki T, którą porównuje się z wartością ${}_{\alpha}T_{(N)}$ odczytywaną z tabel. Hipotezę H_0 (jak zawsze oznaczającą, że nie ma różnicy między obserwacjami) można odrzucić, jeśli

$$T < {}_{\alpha}T_{(N)}$$

Opisane postępowanie jest zupełnie proste z wyjątkiem ustalania rang różnic, które z reguły są rangami wiązanymi, gdyż przy niezbyt rozbudowanych (zazwyczaj) skalach wartości x_i oraz y_i, wartości d_i często się powtarzają i trzeba je opisywać rangami r_i o tej samej wartości, wynikającej z średniej pozycji takich związanych rang w uporządkowanym szeregu wartości. Najlepiej prześledzić to na podanym niżej przykładzie.

9.2.8 Przykład zastosowania testu Wilcoxona

Ocenę skuteczności szczepienia przeciwko określonej chorobie wygodnie jest dokonywać na podstawie ustalenia w surowicy krwi pacjenta tzw. miana przeciwciał. Miano to zwykle jest niezerowe już przed szczepieniem, gdyż pacjent zwykle miewał już wcześniej naturalną styczność z rozważanym antygenem i dlatego ocena skuteczności szczepienia musi być dokonana na podstawie analizy statystycznej. W tabeli 9.4 zestawiono wyniki badań, w których u 13 pacjentów dokonano dwukrotnego oznaczenia miana przeciwciał — przed i po szczepieniu. Badanie ma odpowiedzieć na pytanie o skuteczność szczepienia, zatem zgodnie z przyjętą pragmatyką stawiamy hipotezę H_0, że szczepienie nie miało wpływu na wartość miana przeciwciał.

W tabeli 9.4 wyliczono także różnice d_i dla których trzeba teraz ustalić rangi. Wypisując te różnice w kolejności ich bezwzględnych wartości otrzymujemy następujący szereg liczb:

1 1 2 3 4 5 6 6 6 8 8 10 12
1,5 2 3 4 5 8 10,5 12 13

Poniżej liczb d_i wypisano odpowiadające im rangi r_i. Łatwo zauważyć bez liczenia, że zdecydowanie więcej jest różnic $d_i < 0$ (u większości pacjentów miano przeciwciał wzrasta po szczepieniu), zatem statystykę T wyznaczymy z rang różnic dodatnich:

$$T = 1{,}5 + 4 + 8 = 13{,}5$$

odczytana z tablic wartość krytyczna ${}_{\alpha}T_{(N)} = {}_{0,05}T_{(13)} = 17$, zatem zdecydowanie możemy odrzucić H_0.

Tabela 9.4

Zestaw obserwacji skutków szczepienia podlegających ocenie z wykorzystaniem testu Wilcoxona

Lp.	Miano przeciwciał		różnica	znak	ranga różnicy
	przed szczepieniem	po szczepieniu			
1	63	68	5	–	6
2	41	49	8	–	10,5
3	54	53	1	+	1,5
4	71	75	4	–	5
5	39	49	10	–	12
6	44	41	3	+	4
7	67	75	8	–	10,5
8	56	58	2	–	3
9	46	52	6	–	8
10	37	49	12	–	13
11	61	55	6	+	8
12	68	69	1	–	1,5
13	51	57	6	–	8

Tabele dla testu Wilcoxona zestawiane są głównie dla małych wartości N, ponieważ dla $N > 25$ statystyka T z zadowalającym przybliżeniem może być aproksymowana rozkładem normalnym o parametrach

$$\mu_T = \frac{N(N+1)}{4}$$

$$\sigma_T = \sqrt{\frac{N(N+1)(2N+1)}{24}}$$

i testowanie hipotezy H_0 może odbywać się w oparciu o wartości krytyczne dla standaryzowanego rozkładu normalnego.

9.3 Badanie charakteru rozkładu

9.3.1 Uwagi wprowadzające

W wielu metodach statystycznych możliwość zastosowania określonej metody postępowania warunkowana jest tym, że dane muszą podlegać określonemu, ustalonemu rozkładowi prawdopodobieństwa (na przykład test t-Studenta możliwy jest do zastosowania wyłącznie dla danych o rozkładzie normalnym). Z tego względu zadanie identyfikacji charakteru rozkładu jest w biometrii jednym z ważniejszych i jednym z (niestety) częściej pomijanych.

Przeświadczenie o tym, że dane mają określony rozkład ma z reguły charakter intuicyjny i nieświadomość potrzeby weryfikacji tego założenia jest stałym elementem badań statystycznych. Tymczasem rozbieżność między postulowanym (najczęściej normalnym) a rzeczywistym charakterem rozkładu może być źródłem wyjątkowo przykrych rozczarowań przy stosowaniu metod statystycznych. Dlatego zalecany sposób postępowania musi obejmować weryfikację charakteru rozkładu ilekroć zamierzamy zastosować techniki statystyczne wymagające danych o określonym rozkładzie. W ogólnym wypadku weryfikacja charakteru rozkładu jest zadaniem trudnym właśnie dlatego, że jest to zadanie z samej swojej natury nieparametryczne. W odniesieniu do rozkładów wielowymiarowych problem identyfikacji (w szerokim tego słowa znaczeniu) rozkładu jest jednym z najtrudniejszych zadań statystyki.

Jest jednak zadanie, które jest szczególnie ważne z praktycznego punktu widzenia, a które stosunkowo łatwo rozwiązać. Chodzi o sprawdzenie, czy rozważany zbiór danych spełnia założenie, że dane te zaczerpnięte zostały z populacji o rozkładzie normalnym. Warto podkreślić, że chodzi o jednowymiarowy rozkład normalny, gdyż rozkład wielowymiarowy nie może być weryfikowany tak łatwo. Wspomniane zadanie rozwiązać można z użyciem tak zwanego testu λ-Kołmogorowa, który teraz dokładniej opiszemy.

9.3.2 Test λ–Kołmogorowa

Opisywany test opiera się na następujących zasadach: chcąc stwierdzić, że obserwowane dane podlegają określonemu rozkładowi (najczęściej normalnemu), sporządzamy dystrybuantę (empiryczną) obserwowanego rozkładu analizowanych danych i porównujemy ją z dystrybuantą założonego rozkładu teoretycznego. Podstawą do podjęcia decyzji jest maksymalna wartość różnicy pomiędzy dystrybuantą empiryczną i teoretyczną. Warto zwrócić uwagę, że test ten jest pod wieloma względami podobny do testu Kołmogorowa-Smirnowa, chociaż cel jego stosowania jest w tym wypadku zdecydowanie inny.

Punktem wyjścia do obliczeń jest uporządkowanie ocenianych obserwacji $x_1, ..., x_N$ w ciąg o rosnących wartościach x_i. Dla dalszych obliczeń wygodne jest pogrupowanie

danych w małe przedziały, dla których określa się: prawy koniec każdego przedziału x_i oraz liczbę próbek (obserwacji) zawartych w wybranym przedziale n_i. Jeśli brak innych przesłanek, można rozważane przedziały rozciągać po prostu pomiędzy kolejnymi danymi, przyjmując jako granice przedziału odpowiednio x_{i-1} (wyłącznie) oraz x_i (włącznie) i przyjmując oczywiście $n_i = 1$ ($i = 1, ..., N$). Jako wartość dystrybuanty empirycznej $F_e(x_i)$ przyjąć można w tej sytuacji wartość

$$F_e(x_i) = \frac{\sum_{d \le i} n_j}{N}$$

czyli względne (wydzielone przez łączną liczebność N) skumulowane liczebności klas do x_i włącznie. Przy założeniu określonego typu rozkładu można teraz z wartości (x_1, ..., x_N) wyznaczyć parametry rozkładu (na przykład μ i σ dla rozkładu normalnego), przy czym oczywiście w miarę możności możemy opierać się na zgodnych i efektywnych estymatorach tych parametrów. Na podstawie parametrów rozkładu możemy (z tablic) wyznaczyć teoretyczne wartości dystrybuanty $F_t(x_i)$. Decyzję podejmujemy w oparciu o statystykę λ wyliczaną ze wzoru:

$$\lambda = D\sqrt{N}$$

gdzie

$$D = \sup_{x_i} \left| F_e(x_i) - F_t(x_i) \right|$$

Wyliczona wartość λ porównywana jest z wartością krytyczną ${}_{\alpha}\lambda$ wyznaczoną z tablic. Dla najczęstszego rozważanego przypadku $\alpha = 0{,}05$, ${}_{\alpha}\lambda = 1{,}358$ i nie zależy ani od liczebności N ani od weryfikowanej postaci rozkładu, co jest bardzo wygodne w praktycznych zastosowaniach. Praktyczne wyliczenia z wykorzystaniem testu λ opisane są w kolejnym podrozdziale, ilustrując tezę, że techniczna strona tego testu jest łatwa i prosta w zastosowaniach.

9.3.3 Weryfikacja normalności rozkładu testem λ Kołmogorowa

Jako przykład weryfikacji postaci rozkładu za pomocą testu λ Kołmogorowa rozważymy zagadnienie weryfikacji normalności rozkładu pewnych danych biochemicznych. W tabeli 9.5 zestawiono usystematyzowane wyniki badania 200 próbek osocza krwi, w którym oznaczano ilość tzw. azotu pozabiałkowego („reszta azotowa"). Ponieważ były to próbki krwi ludzi zdrowych (bez uremii), przeto rozrzut obserwowanych wartości był niewielki. Zachodzi pytanie, czy mamy tu do czynienia z rozkładem normalnym.

Tabela 9.5

Wyniki oznaczeń azotu pozabiałkowego (w mg% N) dla 200 próbek osocza krwi

Zawartość azotu	liczba próbek
29,5 – 30,5	12
30,5 – 31,5	23
31,5 – 32,5	35
32,5 – 33,5	62
33,5 – 34,5	44
34,5 – 35,5	18
35,5 – 36,5	6

Aby to sprawdzić dokonano wyliczeń wartości skumulowanych liczebności, a na ich podstawie wyznaczono wielkości $F_e\ (x_i)$ dla wszystkich x_i. Z drugiej strony, poszukując wartości $F_t\ (x_i)$ wyznaczono parametry rozkładu, otrzymując estymatory

$$\mu = 32{,}9$$

$$\sigma = 1{,}4$$

Na tej podstawie wprowadzono zmienną U mającą standaryzowany rozkład normalny

$$u_i = \frac{x_i - \mu}{\sigma}$$

dla tych zmiennych wyznaczono z tabel rozkładu normalnego wartości $F_t\ (x_i)$. Na koniec wyznaczono wartości

$$D_i = |F_e\ (x_i) - F_t\ (x_i)|$$

i zlokalizowano wartość maksymalną D. Wszystkie obliczenia zebrano w tabeli 9.6, gdzie w lewej części zamieszczono wyliczenia zmierzające do określenia $F_e\ (x_i)$ a w prawej — przekształcenia zmierzające do ustalenia $F_t\ (x_i)$.

Ponieważ

$$\lambda = D\sqrt{N} = 0{,}036\sqrt{200} = 0{,}509$$

jest wartością znacznie mniejszą od krytycznej ($_\alpha\lambda = 1{,}358$) przeto nie ma podstaw do odrzucenia hipotezy zerowej H_0, głoszącej, że rozkład jest normalny.

Tablica 9.6

Etapy przekształcania danych z tabeli 9.5 podczas doprowadzenia jej do postaci wymaganej w teście λ-Kołmogorowa

x_i	n_i	n_i skumul.	$F_e(x_i)$	u_i	$F_t(x_i)$	D_i
30,5	12	12	0,060	–1,71	0,044	0,016
31,5	23	35	0,175	–1,00	0,159	0,016
32,5	35	70	0,350	–0,29	0,386	0,036
33,5	62	132	0,660	0,43	0,666	0,006
34,5	44	176	0,880	1,14	0,873	0,007
35,5	18	194	0,970	1,86	0,969	0,001
36,5	6	200	1,000	2,57	0,995	0,005

9.3.4 Weryfikacja charakateru rozkładu za pomocą testu χ^2

Do weryfikacji charakteru rozkładu użyć można także opisanego w rozdziale 6 testu χ^2. Zasada użycia tego testu polega na skonfrontowaniu liczebności empirycznych obserwowanych danych w poszczególnych klasach (wyróżnionych jako przedziały wartości tych danych) z wartościami liczebności teoretycznych, wyznaczonych przy założeniu, że dane podlegają hipotetycznie zakładanemu rozkładowi prawdopdopodobieństwa. Technika przeprowadzania badań jest przy tym następująca:

Przedział zmienności badanych danych dzieli się na r rozłącznych klas i w każdej z tych klas wyznacza się liczebność danych N_i. Oczywiście

$$\sum_{i=1}^{r} N_i = N$$

gdzie N jest liczbą obserwacji. Podział przedziału zmienności na r klas można przeprowadzić dowolnie, najprościej jest jednak zwykle podzielić go równomiernie to znaczy przyjąć taką samą długość każdej z klas. Nie jest to jednak podział optymalny, gdyż wtedy w niektórych przedziałach znajdzie się bardzo dużo obserwacji (duże N_i), a w innych nie będzie ich wcale lub będzie bardzo mało. Jak wiemy z uwag zawartych w rozdziale 6 sytuacja taka jest w teście χ^2 bardzo niekorzystna, gdyż „komórki" w tablicy kontyngencji o małych liczebnościach empirycznych zmuszają do stosowania specjalnych poprawek na nieciągłość testu (na przykład poprawki Yatesa) i należy ich unikać. Tak więc alternatywna propozycja podziału polega na takim wyborze r wyróżnionych klas, by w każdej klasie

mieściła się — w przybliżeniu — taka sama liczba obserwacji N_i. Oczywiście w praktyce badacz wybiera taki podział, jaki jest mu wygodny.

Po dokonaniu podziału określa się prawdopodobieństwa p_i z jakimi, **przy założeniu, że rozkład jest zgodny z hipotetycznie zakładanym,** zmienna losowa przyjmowałaby wartości z poszczególnych wyróżnionych klas. W tym celu oczywiście najpierw trzeba obliczyć (na podstwie całej próby) wartości **parametrów** hipotetycznego rozkładu, a następnie z tablic funkcji gęstości lub dystrybuanty ustalić odpowiednie prawdopodobieństwa teoretyczne p_i. Prawdopodobieństwa te wraz z liczebnością próby N pozwalają na wyznaczenie wartości liczebności teoretycznych w poszczególnych polach tablicy χ^2 i w dalszej konsekwencji na wyznaczenie wartości statystyki χ^2. Statystykę tę można zresztą obliczyć łatwiej ze wzoru

$$\chi^2 = \sum_{i=1}^{r} \frac{(N_i - N p_i)^2}{N p_i}$$

w którym wszystkie oznaczenia zostały wcześniej przedyskutowane. Odrzucenie hipotezy zerowej (głoszącej, że rozważana zmienna ma rozkład zgodny z zadanym rozkładem teoretycznym) następuje w wypadku przekroczenia przez χ^2 wartości krytycznej odczytanej z tablic, jak to dokładnie i szczegółowo opisano w rozdziale 6. Jedynym punktem wymagającym chwili zastanowienia jest kwestia ustalenia liczby stopni swobody. Otóż odczyt z tablic wartości krytycznych powinien nastąpić dla liczby stopni swobody wynoszącej

$$r - k - 1$$

gdzie k jest liczbą parametrów wyznaczanych dla ustalenia rozkładu teoretycznego (np. dla rozkładu normalnego $k = 2$ a dla rozkładu Poissona $k = 1$). Poza tym szczegółem metodyka postępowania jest rutynowo jasna i nie wymaga wyjaśnień, (szczególnie wobec obszernego dyskutowania zasad użycia testu χ^2 w rozdziale 6) dlatego przykłady zostaną tu pominięte.

10. WPROWADZENIE DO WIELOWYMIAROWEJ ANALIZY STATYSTYCZNEJ

10.1 Prezentacja omawianych metod

Każda zaawansowana analiza statystyczna badanych zjawisk wiąże się w istocie z wykrywaniem zależności między rozpatrywanymi zmiennymi, gdyż na gruncie nauk doświadczalnych wyodrębnienie właściwych zmiennych oraz powiązań występujących między nimi jest warunkiem koniecznym skonstruowania jakiejkolwiek teorii.

Testowanie siły oraz rodzaju związku występującego pomiędzy zmiennymi wymaga jednak łącznego (a zarazem równoczesnego) rozpatrywania tych zmiennych — konieczne staje się zatem przejście do wielowymiarowych metod analizy, tzn. do takich w których rozpatruje się wielowymiarowe rozkłady zmiennych losowych i próby pobierane z takich właśnie rozkładów. Metody takie noszą również nazwę wielozmiennowych (zgodnie z ich angielskimi odpowiednikami: multivariate, multidimensional).

Już na pierwszy rzut oka widoczna jest przewaga (z metodologicznego punktu widzenia) metod wielowymiarowych pozwalających na łączne traktowanie wszystkich zmiennych i uwzględnianie zależności typu „wszystko zależy od wszystkiego" nad opisywanymi w pierwszej części skryptu metodami testowania istotności różnic między pojedynczymi zmiennymi.

Rozwój wielowymiarowych metod analizy statystycznej doprowadził do wyodrębnienia dwóch głównych grup pokrywających oddzielne obszary zastosowań:

GRUPA 1 — wykorzystywana jest do wyrażania badanych zmiennych w przestrzeni o mniejszej liczbie wymiarów; wyróżnia się tu:
analizę skupień,
analizę czynnikową i *głównych składowych*.
W przypadku analizy skupień redukcja polega na tworzeniu wspólnych grup zmiennych (skupień), natomiast w przypadku analizy czynnikowej — na tworzeniu nowych wspólnych hipotetycznych zmiennych, tzw. czynników. Analiza skupień wykorzystywana jest również do grupowania obiektów składających się na zbiór danych na podstawie podobieństw wyliczanych między obiektami. Z uwagi na objętościowe ramy skryptu oraz całkowitą odmienność metodologiczną od pozostałych me-

tod analizy wielowymiarowej zdecydowano się na wyłączenie z opisu technik analizy skupień. Proponuje się Czytelnikowi zapoznanie się z literaturą na temat analizy skupień, której wykaz zamieszczony został w dodatku.

GRUPA 2 — służy do badania powiązań między zmiennymi, z rozbiciem ich na zmienne zależne i niezależne. Należy tu zaliczyć:
analizę wariancji wraz z *analizą dyskryminacyjną*,
analizę regresji,
analizę kanoniczną.

Analiza regresji pozwala na łączną ocenę siły oddziaływania na jedną zmienną zależną wielu zmiennych niezależnych. Problem ten można rozszerzyć na badanie siły oddziaływania między dwoma zbiorami zmiennych, tzn. zbiorem zmiennych niezależnych a zbiorem zmiennych zależnych i wówczas ma się do czynienia z analizą kanoniczną. Z uwagi na powiązanie metod obliczeniowych regresji wielokrotnej i regresji krzywoliniowej zdecydowano się na włączenie do tej części skryptu rozdziału opisującego regresję krzywoliniową.

Natomiast analiza wariancji pozwala na: ocenianie zawartości informacyjnej zmiennych losowych jak też ich zbiorów, wykrywanie obszarów wspólnej zmienności wielu zmiennych, wyznaczanie zmiennych redundancyjnych, przeprowadzanie dyskryminacji czy też systematyzowanie nieprzejrzystego zbioru danych. Pewnym problemem jest odpowiednie zaklasyfikowanie analizy dyskryminacyjnej z uwagi na bardzo szerokie pojęciowo znaczenie tego terminu. W skrypcie zdecydowano się zawęzić analizę dyskryminacyjną do pewnego podzbioru metod wielowymiarowej analizy wariancji, rozumiejąc poprzez dyskryminację przyporządkowanie obiektów do jednej z wielu danych klas.

Z góry założone objętościowe ramy skryptu nie pozwoliły na omówienie w nim metod analizy szeregów czasowych. Szeregi czasowe są jednak stosunkowo rzadko wykorzystywane w biometrii (za wyjątkiem prostej analizy trendu), a z drugiej strony istnieje wiele podręczników omawiających ten właśnie zakres materiału[1].

Skuteczne przeprowadzenie dowolnej analizy wielowymiarowej z uwagi na złożoność obliczeniową możliwe jest jedynie przy użyciu komputera. Pociąga to za sobą pewne konsekwencje zarówno w stosunku do prezentacji materiału teoretycznego jak i przykładów. W szczególności nie będą przedstawiane wzory obliczeniowe — z uwagi na konieczność

---

1 Przykładowo można tu polecić książkę: Box G.E.P., Jenkins G.M.: Analiza szeregów czasowych, PWN, Warszawa, 1983

optymalizacji obliczeń różnią się one dość znacznie od podawanych wzorów definicyjnych[2]. Niektóre, dłuższe wyprowadzenia zastąpione zostały przedstawieniem samej idei oraz schematu postępowania. Ograniczeniu uległa — w stosunku do pierwszej części skryptu — liczba przykładów, gdyż czasami ich prezentacja bez pomocy komputera musiałaby ograniczyć się jedynie do trywialnych.

Warto jednak wspomnieć, że wszystkie omawiane w skrypcie metody wielowymiarowe mają swoje implementacje programowe. W przypadku mikrokomputerów typu IBM PC polecić można takie zintegrowane pakiety programowe jak: SPSS/PC+, StatGraphics, SYSTAT/SYGRAPH czy też BMDP. Wszystkie te pakiety pozwalają również na graficzną prezentację zarówno danych jak i wyników obliczeń, co w przypadku metod wielowymiarowych ma bardzo istotne znaczenie.

Opisywane w tej części skryptu metody analizy wymagają zarówno znajomości wielowymiarowego rozkładu normalnego i rozkładów z niego wyprowadzanych, jak i znajomości rachunku macierzowego. Pewne niezbędne informacje na ten temat przedstawione zostały (z konieczności skrótowo) w dodatkach. Zakłada się jednak znajomość przez czytelnika podstaw rachunku macierzowego i analizy matematycznej. Informacje zawarte w tej części skryptu są niezależne od części pierwszej w tym sensie, że mogą być czytane niezależnie od niej, jednak pod warunkiem znajomości podstawowego kursu statystyki matematycznej. Również poszczególne rozdziały stanowią odrębne całości.

Metody wielowymiarowej analizy statystycznej posiadają już dość bogatą literaturę. Na końcu skryptu zamieszczono odpowiedni wykaz literatury starając się dobierać pozycje nie tylko łatwiej dostępne dla polskiego czytelnika, lecz również w najlepszy sposób poszerzające przedstawione w skrypcie rozważania.

10.2 Obiekty i cechy w analizie wielowymiarowej

Rozpoczęcie jakichkolwiek analiz statystycznych wymaga uprzedniego ustalenia zbioru tzw. jednostek statystycznych, nazywanych również obiektami, oraz zbioru opisujących badane zjawiska cech, nazywanych zmiennymi.

W każdym badaniu statystycznym występują zatem dwa rodzaje podstawowych wielkości, a mianowicie zbiór obiektów:

$$I = \{I_1, I_2, \ldots, I_M\}$$

---

2 Po raz pierwszy z tym problemem czytelnik zetknął się zapewne czytając o sposobie liczenia wariancji zmiennej losowej — w przypadku metod wielowymiarowych i rachunku macierzowego różnica między odpowiednimi wzorami jest olbrzymia.

gdzie M jest liczbą badanych obiektów, oraz zbiór cech:

$$Z = \{Z_1, Z_2, \ldots, Z_N\}$$

gdzie N jest liczbą rozpatrywanych mierzalnych cech. Zakłada się, że zbiór cech Z charakteryzuje każdy obiekt ze zbioru obiektów I.

Proces badawczy nie jest oczywiście prowadzony bezpośrednio na obiektach lub cechach, lecz na realizacjach cech. Rezultat pomiaru k-tej cechy obiektu I_i oznacza się jako x_{ik}, zatem wektor

$$\boldsymbol{X}_i = (x_{i1}, x_{i2}, \ldots, x_{iN}) \tag{10.1}$$

odpowiada pomiarowi wszystkich cech i-tego obiektu.
Odpowiednikiem zbioru wektorów pomiarów:

$$\hat{\boldsymbol{X}} = \{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_M\}$$

opisujących obiekty ze zbioru I jest **macierz danych** $\boldsymbol{X}$ zdefiniowana jak poniżej:

$$\boldsymbol{X} = (x_{ik}), \qquad i = 1, 2, \ldots, M \qquad k = 1, 2, \ldots, N \tag{10.2}$$

W wierszach macierzy $\boldsymbol{X}$ występują obiekty $\boldsymbol{X}_i$, które traktuje się jak punkty lub wektory usytuowane w N-wymiarowej przestrzeni nazywanej przestrzenią obiektów. W przestrzeni tej każdą z N cech opisujących obiekty $\boldsymbol{X}_i$ przedstawia się w postaci osi współrzędnych; każdy obiekt $\boldsymbol{X}_i$ jest wtedy punktem lub zaczepionym w początku układu wektorem tej przestrzeni o współrzędnych określonych wzorem (10.1).

Należy zauważyć, że w ogólnym przypadku układ N współrzędnych nie musi być układem prostokątnym (z uwagi na współzależność cech), nie należy również utożsamiać odległości między punktami z odległością euklidesową.

Analogicznie kolumny macierzy danych $\boldsymbol{X}$ są realizacjami cech Z_j, które również traktuje się odpowiednio jako punkty lub wektory:

$$\boldsymbol{Y}_j = \left(x_{1j}, x_{2j}, \ldots, x_{Mj}\right)^T \tag{10.3}$$

gdzie $\boldsymbol{Y}_j$ jest wektorem realizacji cechy Z_j, $j = 1, 2, \ldots, N$. Punkty te są usytuowane w M-wymiarowej przestrzeni cech. W przestrzeni tej każda cecha Y_j jest zatem traktowana jako punkt, albo jako wektor skierowany od początku układu współrzędnych do danego punktu; współrzędne tego punktu określone są wzorem (10.3). Osiami układu współrzędnych tej przestrzeni są obiekty, liczba osi wynosi zatem M.

Cechy występujące w macierzy danych posiadają różne jednostki miar, również ich wartości bezwzględne różnią się znacznie między sobą. Aby móc rozpatrywać je wszystkie łącznie, trzeba przeprowadzić procedurę unifikacji, tzn. uwolnienia od jednostek miary i ustalenia jednakowego zakresu zmienności.

Najczęściej stosowany sposób uwalniania cech od jednostek miary to standaryzacja przekształcająca wartość każdej ze zmiennych w poniższy sposób:

$$y_{mn} = \frac{x_{mn} - \bar{x}_n}{\sigma_n}, \qquad m = 1, 2, ..., M \qquad n = 1, 2, ..., N \tag{10.4}$$

gdzie:

x_{mn} — m-ta realizacja n-tej zmiennej,
$\bar{x}_n$ — wartość oczekiwana (średnia) n-tej zmiennej,
σ_n — odchylenie standardowe n-tej zmiennej.

Inny sposób, transformujący wartości każdej zmiennej do przedziału domkniętego <0, 1> wyraża się wzorem:

$$x_{mn} = \frac{x_{mn} - \min\limits_{m} \{x_{mn}\}}{\max\limits_{m} \{x_{mn}\} - \min\limits_{m} \{x_{mn}\}} \tag{10.5}$$

Czasami stosuje się bardziej wyrafinowane metody unifikacji. Traktując realizacje wielowymiarowych zmiennych losowych jako punkty (lub wektory) usytuowane w pewnej wielowymiarowej przestrzeni, można zauważyć, że są one zróżnicowane pod względem poziomu wartości opisujących je zmiennych, to jest proporcji poziomów ich wartości. Wówczas o strukturze wartości zmiennych informuje kąt między rozpatrywanymi wektorami, a o poziomie wartości zmiennych — długość poszczególnych wektorów.

Możliwe jest zastosowanie takiej unifikacji, aby otrzymać dane jednorodne albo pod względem struktury, albo pod względem poziomu wartości zmiennych.

Przykładowo przekształcenie, które z wartości każdej zmiennej eliminuje składnik struktury, a pozostawia składnik poziomu, określa się następująco:

niech $\boldsymbol{X}_n = \{x_{1n}, x_{2n}, ..., x_{mn}, ..., x_{Mn}\}^T$ będzie wektorem realizacji n-tej cechy; wówczas dane transformowane

$\boldsymbol{V}_n = \{v_{1n}, v_{2n}, ..., v_{mn}, ..., v_{Mn}\}^T$ wylicza się korzystając na podstawie wzoru

$$V_{mn} = \frac{x_{mn}}{\sum\limits_{m=1}^{M} x_{mn}}, \quad \text{gdzie } n = 1, 2, ..., N \tag{10.6}$$

Tak przekształcone zmienne charakteryzują się tym, że ich długości (rozumiane oczywiście jako długości wektorów) są jednostkowe. Dla wszystkich zmiennych V_n spełniony jest zatem warunek

$$\|V_n\| = \sum_{m=1}^{M} v_{mn} = 1 \ , \qquad n = 1, 2, ..., N$$

podczas gdy zmienne X_n mają zazwyczaj różne długości. Ze zmiennych V_n wyeliminowano w ten sposób jednostki miary i doprowadzono do ustalenia długości zmiennych. Tak więc obiekty (dane statystyczne) będące punktami przekształconej przestrzeni obiektów, są opisywane przez wartości v_{mn} cech, które mają tylko składnik reprezentujący poziom wartości cech, tj. składnik obrazujący *potencjał*, *skalę*, *pozycję* czy też *rangę* danego obiektu w rozpatrywanym zbiorze obiektów. Wartości v_{mn} są zatem udziałem wartości n-tej zmiennej m-tego obiektu w sumie wartości tej samej zmiennej we wszystkich obiektach zbioru.

Inne z kolei przekształcenie tak transformuje zbiór obiektów, aby jego elementy w wielowymiarowej przestrzeni cech utworzyły smugę punktów o kształcie elipsoidy. Do takiego zbioru punktów (o ile jest on dodatkowo spójny) zastosować można model liniowy, wymagający normalnego rozkładu[3] wielowymiarowej zmiennej losowej (o rozrzucie danych właśnie w postaci elipsoidy). Obiekty spełniające powyższy warunek mają wtedy zbliżoną strukturę wartości zmiennych, tzn. małe zróżnicowanie proporcji wartości odpowiednich zmiennych.

Przekształcenie eliminujące oddziaływanie poziomu wartości zmiennych, a pozostawiające wpływ struktury wartości tych zmiennych przeprowadza się następująco:

niech $P_m = \{x_{m1}, x_{m2}, ..., x_{mM}\}^T$ będzie wektorem realizacji m-tego obiektu; wówczas dane transformowane

$$Z_m = \{z_{m1}, z_{m2}, ..., z_{mN}\}$$

wylicza się jako

$$z_{mn} = \frac{\dfrac{x_{mn}}{\sum_{m=1}^{M} x_{mn}}}{\sum_{n=1}^{N} \dfrac{x_{mn}}{\sum_{m=1}^{M} x_{mn}}} \qquad (10.7)$$

3 Informacje na temat wielowymiarowego rozkładu normalnego zawarte są w jednym z dodatków

Dowodzi się wówczas, że najmniejsza wartość odległości między obiektami (których współrzędne przekształcono zgodnie ze wzorem (10.7)) występuje wtedy, gdy reprezentujące te obiekty wektory są równoległe, tj. gdy mają identyczną strukturę wartości zmiennych. Natomiast największa wartość odległości między obiektami ma miejsce przy występowaniu wektorów prostopadłych, czyli obiektów o maksymalnie zróżnicowanej strukturze wartości zmiennych. Ogólnie można stwierdzić, że im obiekty mniej różnią się co do struktury wartości zmiennych, tym odległość między obiektami jest mniejsza.

Wprowadzenie macierzy danych a następnie transformowania zmiennych implikuje również wyliczanie odległości między obiektami (lub też zmiennymi). W sposób niejawny pod pojęciem odległości rozumie się zazwyczaj odległość euklidesową lub tzw. *miejską*, rozumianą jako suma wartości bezwzględnych różnic poszczególnych składowych. Nie zawsze jednak stosuje się tylko te dwie metryki. Często dobór metryki związany jest z obszarem zastosowań danej procedury badawczej (przykładowo: jak obliczyć „odległość" między dwoma łańcuchami kwasu DNA lub jak zmierzyć „odległość" między wynikami testu psychologicznego).

11. WIELOWYMIAROWA ANALIZA WARIANCJI I ANALIZA DYSKRYMINACYJNA

Przypomnijmy pokrótce, że analiza wariancji zajmuje się badaniem związku między pewnymi czynnikami zewnętrznymi wpływającymi na wartości pomiarowe zmiennych losowych a tymi zmiennymi losowymi. Weryfikuje się przy tym, czy pomierzone cechy w poszczególnych obiektach, będących pod działaniem różnych wpływów zewnętrznych, wykazują zmienność.

W przypadku znanej nam już jednowymiarowej analizy wariancyjnej uwzględnia się tylko jedną cechę, natomiast w przypadku wielowymiarowej analizy wariancyjnej rozpatruje się cały ich szereg — wektor cech. Poszczególne cechy składowe (komponenty) wektora cech są na ogół wzajemnie od siebie statystycznie zależne, podczas gdy wektory cech opisujące różne obiekty muszą być wzajemnie statystycznie niezależne.

Analiza wariancji zakłada, że w odniesieniu do badanych cech rozpatrywane zmienne są ciągłe. Ponadto przyjmuje się jeszcze, że cechy lub wektory cech mają rozkład normalny. Aczkolwiek w praktyce warunki te nie zawsze są spełniane, mimo to analiza wariancji może być z pożytkiem stosowana — będziemy jeszcze dyskutować założenia analizy wariancji i możliwe do przyjęcia odchylenia od nich.

W przeciwieństwie do cech rozpatrywane w analizie wariancyjnej czynniki zewnętrzne mają charakter jakościowy. Do każdego takiego czynnika należy pewna liczba różnych stanów, które zwykle nazywa się poziomami. Zakłada się, że każdy obiekt odnośnie do każdego występującego czynnika może być przyporządkowany dokładnie jednemu poziomowi.

W zależności od liczby rozpatrywanych czynników rozróżnia się analizy wariancyjne jednoczynnikowe i wieloczynnikowe. W wieloczynnikowej analizie wariancji uwzględnia się jednocześnie kilka czynników, przy czym każdy obiekt włączony jest w schemat komórek, które powstają przez kombinacje poziomów różnych czynników. Jeśli każda z istniejących komórek jest obsadzona przez obiekty, to mamy wówczas do czynienia z eksperymentem o klasyfikacji krzyżowej. Jeśli zaś każdy poziom pewnego czynnika B może wystąpić jedynie w kombinacji ze ściśle określonym poziomem czynnika A, to wtedy mówimy o tzw. klasyfikacji hierarchicznej. Obok tzw. efektów głównych poszczególnych czynników rozpatruje się w analizie wariancyjnej wieloczynnikowej jeszcze tzw. efekty interakcji pomiędzy czynnikami.

Rozpatrując zatem dobrze znany model liniowy zależności między cechami

$$Y_i = \sum_{j=1}^{k_i} \beta_{ij} X_{ij} + \beta_{i0} + \varepsilon_i \qquad \text{dla } i = 1, ..., p$$

(gdzie Y_i to zmienne zależne, tzn. objaśniane przez model, natomiast X_{ij} stanowią zbiór zmiennych niezależnych, a ich liczbę dla każdego i-tego równania opisuje indeks k_i) dochodzimy do wniosku, że będzie on opisywał analizę wariancji wtedy, gdy zmienne niezależne X_{ij} będą przybierać jedynie wartości 0 lub 1, a zatem będą charakteryzować one pewien efekt jako występujący bądź nie[1]. Rozważając wspomniany liniowy model analizy wariancji dla ustalonego i, tzn. rozważając dane równanie niezależnie od pozostałych, mamy do czynienia z jednowymiarową analizą wariancji. Przypadek wielowymiarowy polega na łącznym rozpatrywaniu wszystkich równań.

Jeśli porównamy p-wymiarową analizę wariancji z p jednowymiarowymi analizami wariancyjnymi, które mogą być przeprowadzone dla p różnych cech, to dojdziemy do wniosku, że:

1. Jednowymiarowe analizy wariancyjne dostarczają nam specjalnych informacji odnośnie do p poszczególnych cech, przy czym nie znajdują tu odbicia powiązania (zależności) między cechami. Od cechy do cechy otrzymuje się tu inne wyniki; jedna cecha daje bardziej istotne różnice między poziomami, inna znów mniej istotne. Dlatego też jakaś zbiorcza ocena działania wszystkich cech na raz nie jest tu jeszcze możliwa. W przypadku zaś wielowymiarowej analizy wariancyjnej otrzymywane wyniki opierają się na całości wszystkich rozpatrywanych cech, przy czym grają tu niepoślednią rolę korelacje między nimi. Wielowymiarowa analiza wariancji umożliwia zatem uzyskanie pełnego poglądu co do wzajemnych związków ukrytych w wielowymiarowym materiale danych.
2. Wyniki uzyskane na drodze p-wymiarowej analizy wariancyjnej nie dają się sprowadzić do wyników uzyskanych z p jednowymiarowych analiz wariancyjnych. Nie jest mianowicie tak, że zawsze te cechy, które w jednowymiarowym teście są najbardziej efektywne, również i w wielowymiarowym teście, odnoszącym się do zbioru tych cech jako całości, mają podobne właściwości. Może się tu nawet zdarzyć, że cechy, które badane osobno nie wykazują żadnych istotności i stąd najczęściej ignorowane są w tradycyjnej jednozmiennej analizie wariancyjnej, w ich wielowymiarowym połączeniu wykazują bardzo dużą moc informacyjną.

---

1 Jeśli wszystkie zmienne niezależne są mierzalne, to mamy do czynienia z analizą regresji, natomiast jeśli niektóre tylko spośród X_{ij} są niemierzalne a pozostałe są mierzalne, wówczas stosujemy nie omawianą w tym skrypcie analizę kowariancji

Zatem wielowymiarowa analiza wariancji przynosi w porównaniu z jednowymiarową analizą wariancyjną rzeczywiste wzbogacenie pojęciowe. Ponadto wielowymiarowa analiza wariancji stanowi bazę metodyczną dla innych metod statystycznych, zwłaszcza w powiązaniu z analizą dyskryminacyjną[2]. Za pomocą tych metod można oceniać zawartość informacyjną zmiennych losowych, jak też ich zbiorów, można wyznaczać zmienne redundancyjne, systematyzować nieprzejrzysty zbiór danych, przeprowadzać dyskryminację danych obiektów. Możemy wreszcie za pomocą pewnej transformacji przejść do odpowiednio szczuplejszego zbioru cech o możliwie wysokiej zawartości informacyjnej (cechy dyskryminacyjne), gdzie uzyskamy przejrzysty obraz współzależności eksperymentalnych w przestrzeni o niższym wymiarze.

Materiał przedstawiany w tym rozdziale podzielimy na trzy podrozdziały, osobno rozważać będziemy wielowymiarowy przypadek jednej lub dwóch populacji, w drugim podrozdziale uogólnimy te rozważania na wielowymiarowy przypadek klasyfikacji pojedynczej przy większej niż dwie liczbie populacji. Na koniec wreszcie, wspomnimy o wielowymiarowym przypadku klasyfikacji wielokrotnej.

11.1 Wielowymiarowa analiza wariancji w przypadku jednej lub dwóch populacji

11.1.1 Oceny wektora średnich populacji i macierzy kowariancji w łącznym rozkładzie normalnym

Założymy, że dla każdego obiektu dana jest p-wymiarowa zmienna losowa, którą stanowi ciąg p wartości pomiarowych, odpowiadających określonym cechom tych obiektów:

$$y = [y_1, y_2, ..., y_p]^T$$

W celu analizy statystycznej przyjmiemy dalej, że wektory obserwacji rozważanych obiektów tworzą p-wymiarową populację o rozkładzie normalnym $N(\mu, \Sigma)$. Wielkość

$$\mu = [\mu_1, \mu_2, ..., \mu_p]^T$$

---

2 Przez dyskryminację rozumieć będziemy w tym rozdziale procedurę przyporządkowania obiektów do jednej z wielu danych klas, innymi słowy procedurę różnicowania obiektów.

oznacza tu wektor wartości średnich w populacji, a macierz

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1p} \\ \dots & \dots & \dots & \dots \\ \sigma_{p1} & \sigma_{p2} & \dots & \sigma_{pp} \end{bmatrix}$$

jest macierzą kowariancji. Oczywiście zarówno μ jak i Σ nie są znane.

W przypadku, gdy mamy próbę złożoną z n wektorów wyników obserwacji

$$y_1 = [y_{11}, y_{21}, \dots y_{p1}]^T,$$
$$y_2 = [y_{12}, y_{22}, \dots y_{p2}]^T,$$
$$\dots\dots\dots\dots\dots$$
$$y_n = [y_{1n}, y_{2n}, \dots y_{pn}]^T,$$

a więc gdy znane są z obserwacji wartości pomiarowe p cech n obiektów ($n \geq 2$), wówczas można uzyskać nieobciążone estymatory wielkości μ i Σ. Oceną wektora μ jest wektor

$$y_\cdot = \frac{1}{n} \sum_{j=1}^{n} y_j \tag{11.1}$$

a oceną macierzy Σ jest macierz

$$S = \frac{1}{n-1} \sum_{j=1}^{n} (y_j - y_\cdot)(y_j - y_\cdot)^T = \frac{1}{n-1} \left(\sum_{j=1}^{n} y_j y_j^T - n y_\cdot y_\cdot^T \right) \tag{11.2}$$

Macierz S jest półokreślona dodatnio, a w większości przypadków praktycznych jest to nawet dodatnio określona macierz symetryczna. Na głównej przekątnej macierzy S znajdują się wariancje poszczególnych cech. Współczynniki korelacji r_{ij} dwóch cech y_i i y_j otrzymuje się z równania:

$$r_{ij} = \frac{s_{ij}}{\sqrt{s_{ii}\, s_{jj}}}$$

Oszacowanie $y_\cdot$ ma rozkład $N(\mu, \Sigma/n)$, natomiast macierz S podlega rozkładowi Wisharta[3] $W(\Sigma/(n\text{-}1), n\text{-}1)$.

---

3 patrz dodatek 3

Jeśli mamy do czynienia z n wektorami wyników obserwacji $y_1, y_2, ..., y_n$ $(n \geq p + 1)$, to odpowiadający im model liniowy ma postać:

$$y_i = \mu + \varepsilon_j \qquad (j = 1, ..., n) \tag{11.3}$$

a jako hipotezę zerową przyjmiemy na początek:

$$H_0 : \mu = 0 \tag{11.4}$$

I w założeniach analizy wariancji i w postaci hipotezy zerowej występują analogie z analizą jednowymiarową. Fakt ten musi znaleźć odbicie w postaci statystyki testującej hipotezę zerową. Przypomnijmy sobie zatem jak skonstruowany jest test F w przypadku analizy jednowymiarowej i np. dwuczynnikowej. Jest on ilorazem dwóch wariancji, z założenia większej przez mniejszą. Wariancja mniejsza, mianownik tego ilorazu, to wariancja zmiennej Y nie wyjaśniona ani przez wpływ czynnika A, ani przez wpływ czynnika B, ani przez wpływ interakcji obu czynników. Stanowi ją naturalny rozrzut pomiarów wokół średniej w każdej z podpróbek próby losowej (czyli w komórkach). To, co wstawia się do mianownika testu F, jest uśrednioną wariancją z wariancji tych podpróbek.

Przez analogię, w przypadku wielowymiarowym takie ważone wariancje tworzy się równocześnie dla p zmiennych, według zasady obowiązującej dla przypadku jednowymiarowego. Mianowicie rozrzut wyników w każdej komórce jest zależny nie tylko od wariancji wyliczonej dla każdej ze zmiennych $y_1, y_2, ..., y_p$ oddzielnie, ale także od kowariancji między tymi zmiennymi. Tak więc dla każdej komórki mamy jedną macierz kowariancji. Macierz ważoną z wielu podprób tworzy się przez „uśrednienie" tych wszystkich macierzy. W przypadku jednowymiarowym mieliśmy średnią wariancję, teraz jest to średnia macierz kowariancji S. Macierz S „częściowa", tzn. jeszcze nie podzielona przez stopnie swobody oznaczana jest tradycyjnie przez G.

Podobnie jest z licznikiem testu F. Postać licznika w analizie jednowymiarowej zależy od tego, która hipoteza zerowa jest weryfikowana. Wartość licznika jest także wariancją i ma interpretację wariancji wyjaśnionej: albo przez wpływ czynnika A na Y, wtedy wariancja wyznaczona jest przez rozrzut średnich z poziomu tego czynnika wokół średniej globalnej (z całej próby), albo przez wpływ czynnika B na Y, wtedy wariancja wyznaczona jest przez rozrzut średnich z poziomu czynnika B wokół średniej globalnej, albo przez wpływ interakcji, wtedy wariancja w liczniku wyznaczona jest przez rozrzut interakcji. Konstruując identyczne rozrzuty średnich z poziomów czynnika przy p zmiennych zamiast jednej liczby (wariancji wyjaśnionej) otrzymuje się macierz, której elementy zależne są wyłącznie od tych rozrzutów. Macierz tę w formie nieco wcześniejszej, tzn. nie podzieloną przez liczbę stopni swobody oznacza się literą H.

Wprowadźmy zatem teraz opisane nowe macierze, a mianowicie

$$H = ny.y.^T \tag{11.5}$$

oraz

$$G = \sum_{j=1}^{n} (y_j - y.)\,(y_j - y.)^T = \sum_{j=1}^{n} y_j y_j^T - ny.\,y.^T \tag{11.6}$$

Macierze te noszą również własne nazwy: G — macierz błędu, H — macierz obiektowa.

Oszacowanie mianownika testu to macierz G, zatem w teście wystąpi macierz G^{-1}, a ściślej symetryczna macierz HG^{-1}. Do statystyki testowej wstawimy oczywiście nie macierz HG^{-1}, lecz pewną wyliczoną z niej wartość. Z postulatów formalnych dotyczących postaci testu weryfikującego H_0 (których tu nie będziemy analizować) wynika, że test musi być funkcją wartości własnych macierzy HG^{-1}. Może zatem istnieć wiele rozwiązań tego problemu. Najczęściej używaną funkcją jest kryterium śladowe, zwane inaczej statystyką T^2 Hotellinga:

$$T^2 = tr\,(HG^{-1})$$

Zmienna losowa T^2 Hotellinga jest uogólnieniem zmiennej losowej t Studenta opartej na jednej zmiennej.

Rozkład z próby statystyki T^2 można aproksymować za pomocą rozkładu F. Dla całej statystyki testowej $\tilde{F}$, przy prawdziwości hipotezy zerowej, wyznacza się stopnie swobody ν_1 i ν_2, przy których $\tilde{F}$ ma rozkład w przybliżeniu zgodny z rozkładem F Snedecora[4].

Zgodnie z przeprowadzonymi powyżej rozważaniami jako statystykę testową przyjmujemy

$$\tilde{F} = \frac{f_2 - p + 1}{f_1\, p}\, tr\,(HG^{-1}), \qquad f_1 = 1,\ f_2 = n - 1 \tag{11.7}$$

Statystyka $\tilde{F}$ ma przy założeniu prawdziwości hipotezy zerowej H_0 dokładnie rozkład F ze stopniami swobody

$$\nu_1 = p,\ \ \nu_2 = f_2 - p + 1 \tag{11.8}$$

W konsekwencji więc przy poziomie istotności α postawioną hipotezę odrzucamy, gdy

---

4 dla zaznaczenia tego przybliżenia nad literą F znajduje się wężyk

$$\tilde{F} > F_{p, f_2 - p + 1, \alpha}$$

Po odpowiednich przeliczeniach otrzymujemy ostateczny wzór na wartość statystyki testowej

$$\tilde{F} = \frac{(n-p)\,n}{(n-p)\,p}\; y.^T S^{-1}\, y. \tag{11.9}$$

gdzie stopniami swobody są $\nu_1 = p, \quad \nu_2 = n - p.$
Przejście do ogólniejszej hipotezy zerowej

$$H_0 : \mu = \mu^* \tag{11.10}$$

otrzymuje się przez niewielką korektę statystyki danej wzorem (9), a mianowicie:

$$\tilde{F} = \frac{(n-p)\,n}{(n-p)\,p} (y. - \mu^*)^T S^{-1} (y. - \mu^*) \tag{11.11}$$

Stopniami swobody są identycznie jak poprzednio: $\nu_1 = p$ oraz $\nu_2 = n - p$.

Przykład 1.

Rozważmy dane dotyczące pewnej grupy noworodków. Na 20 noworodkach dokonano pomiarów wagi oraz długości ciała otrzymując w wyniku wektor wartości średnich

$$y. = \begin{bmatrix} 3509 \\ 51{,}5 \end{bmatrix}$$

gdzie liczba w pierwszym wierszu oznacza średnią wagę noworodka w gramach, a liczba w drugim wierszu — długość jego ciała w centymetrach. Oszacowanie macierzy kowariancji jest następujące

$$S = \begin{bmatrix} 213683 & 823 \\ 823 & 4{,}89 \end{bmatrix}$$

Otrzymujemy stąd na odchylenia standardowe rozważanych dwóch cech noworodka wartości

$$\sigma_{11} = 462, \quad \sigma_{22} = 2{,}21,$$

a na współczynnik korelacji obu tych wielkości wartość

$$r_{12} = 0{,}804.$$

Jeśli chcemy sprawdzić, czy prawdziwa średnia waga noworodka wynosi 3000 g i czy prawdziwa średnia długość jego ciała wynosi 50 cm, tzn. gdy zgodnie z wzorem (11.10) testujemy hipotezę w której

$$\mu^* = \begin{bmatrix} 3000 \\ 50 \end{bmatrix},$$

to zgodnie z (11.11) otrzymujemy

$$\tilde{F} = 12.64, \qquad \nu_1 = 2, \qquad \nu_2 = 18.$$

Ponieważ $F_{2,18,0.05} = 3.55$, więc postawioną hipotezę zerową odrzucamy. ■

Z weryfikacją hipotezy zerowej $H_0 : \mu = \mu^*$ wiąże się problem wyznaczania obszaru ufności B dla wektora wartości średnich μ. Obszar ufności B przy danej n-obiektowej próbie należy tak określić, żeby zawierał on faktyczny wektor wartości średnich μ przy z góry zadanym prawdopodobieństwie wynoszącym $1 - \alpha$. Żądanie to będzie spełnione, gdy do obszaru B zaliczać będziemy każdy wektor m, spełniający warunek

$$\tilde{F} = \frac{(n-p)\,n}{(n-p)\,p} (y. - m)^T S^{-1} (y. - m) \leq F_{p,\,n-p,\,\alpha} \tag{11.12}$$

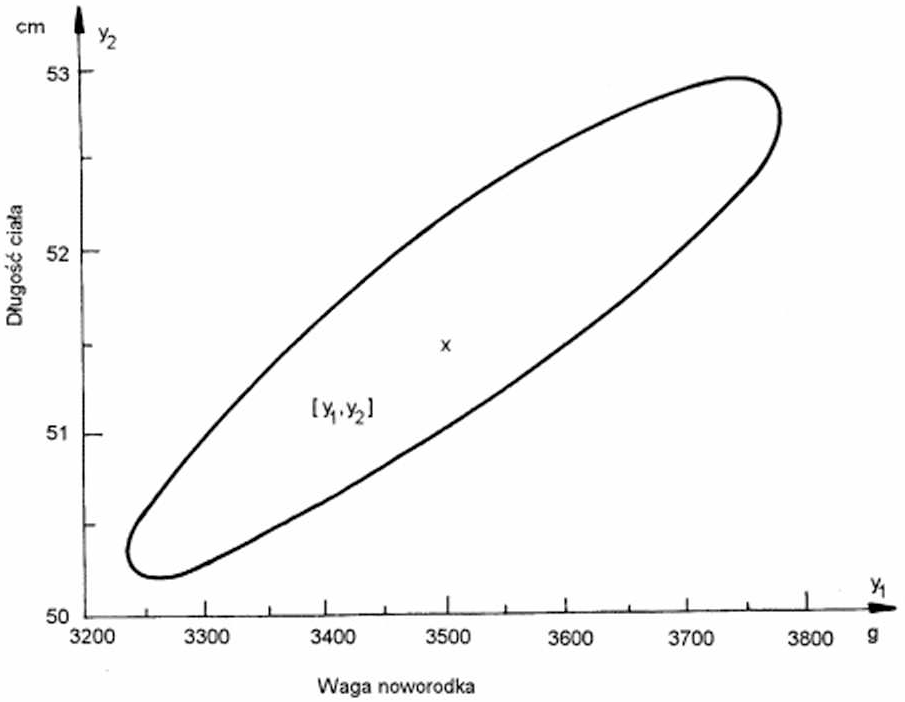
Przy danej próbie obszar B obejmuje punkty położone wewnątrz p-wymiarowej elipsoidy, której środkiem jest akurat $y.$.

Przykład 2.

Korzystając z danych liczbowych przykładu 1 otrzymujemy na obszar ufności wektora średnich $\mu = [\mu_1, \mu_2]^T$ przy poziomie ufności 0.95 następującą nierówność

$$[m_1 - 3509, m_2 - 51{,}5] \begin{bmatrix} 0{,}000126 & -0{,}0211 \\ -0{,}0211 & 5{,}48 \end{bmatrix} \begin{bmatrix} m_1 - 3509 \\ m_2 - 51{,}5 \end{bmatrix} \leq 3{,}55$$

Obszar ten jest pokazany na rys. 11.1. Liczby m_1 (waga noworodka) i m_2 (długość jego ciała) wchodzą w rachubę jako prawdziwe wartości średnie badanego zbioru noworodków, gdy spełniają one (tj. m_1 i m_2) ostatnią nierówność, tzn. gdy punkt (m_1, m_2) leży wewnątrz narysowanej elipsoidy. ■



Rys. 11.1 Obszar ufności dla wektora wartości średnich przy poziomie ufności $1 - \alpha = 0{,}95$

11.1.2 Różnica dwóch wektorów średnich przy nieznanej macierzy kowariancji

Rozważamy dwie populacje obiektów. Zakładamy, że p-wymiarowe wektory wyników obserwacji w pierwszej populacji mają rozkład $N(\mu_1, \Sigma)$, natomiast w drugiej populacji — rozkład $N(\mu_2, \Sigma)$. Oznacza to, że obie populacje pokrywają się jeśli chodzi o macierze kowariancji, podczas gdy wektory wartości średnich mogą się różnić. Zarówno μ_1 jak i μ_2 oraz Σ nie są znane.

Wyjdziemy od dwóch prób, wziętych z obu tych populacji. Niech zatem wektorami wyników pomiarowych pierwszej próby będą $y_{11}, y_{12}, ..., y_{1n_1}$, a wektorami drugiej klasy $y_{21}, y_{22}, ..., y_{2n_2}$ ($n_1 \geq 1$, $n_2 \geq 1$, $n_1 + n_2 \geq p + 2$). Wówczas równania rozważanego modelu mają postać

$$y_{1k} = \mu_1 + \varepsilon_{1k} \qquad (k = 1, ..., n_1), \tag{11.13}$$

$$y_{2k} = \mu_2 + \varepsilon_{2k} \qquad (k = 1, ..., n_2). \tag{11.14}$$

Hipotezą do weryfikacji jest hipoteza

$$H_0 : \mu_1 - \mu_2 = 0 \tag{11.15}$$

Dla każdej próby obliczamy wektor wartości średnich

$$y_{1.} = \frac{1}{n_1} \sum_{k=1}^{n_1} y_{1k}, \qquad y_{2.} = \frac{1}{n_2} \sum_{k=1}^{n_2} y_{2k}$$

i wyznaczamy macierz kowariancji S, która daje nam ocenę prawdziwej wewnątrzgrupowej macierzy kowariancyjnej Σ:

$$S = \frac{1}{n_1 + n_2 - 2} \left(\sum_{k=1}^{n_1} (y_{1k} - y_{1.})(y_{1k} - y_{1.})^T + \sum_{k=1}^{n_2} (y_{2k} - y_{2.})(y_{2k} - y_{2.})^T \right).$$

Wprowadzamy analogicznie jak poprzednio macierze G i H:

$$H = \frac{n_1 n_2}{n_1 + n_2} (y_{1.} - y_{2.})(y_{1.} - y_{2.})^T, \tag{11.16}$$

$$G = \sum_{k=1}^{n_1} (y_{1k} - y_{1.})(y_{1k} - y_{1.})^T + \sum_{k=1}^{n_2} (y_{2k} - y_{2.})(y_{2k} - y_{2.})^T. \tag{11.17}$$

Jeśli zatem uwzględnimy równość

$$S = \frac{1}{n_1 + n_2 - 2} G \tag{11.18}$$

to zgodnie ze wzorem (11.7) jako statystykę testową otrzymamy wielkość

$$\tilde{F} = \frac{n_1 + n_2 - p - 1}{p\,(n_1 + n_2 - 2)} \frac{n_1 n_2}{n_1 + n_2} tr\left((y_{1.} - y_{2.})(y_{1.} - y_{2.})^T S^{-1} \right)$$

czyli

$$\tilde{F} = \frac{n_1 + n_2 - p - 1}{p\,(n_1 + n_2 - 2)} \frac{n_1 n_2}{n_1 + n_2} (y_{1.} - y_{2.})^T S^{-1} (y_{1.} - y_{2.}) \tag{11.19}$$

ze stopniami swobody $\nu_1 = p$, $\nu_2 = n_1 + n_2 - p - 1$.

Przy prawdziwości hipotezy H_0 statystyka $\tilde{F}$ ma dokładnie rozkład F. Hipotezę H_0 odrzucimy, jeżeli okaże się, że $\tilde{F} > F_{\nu_1, \nu_2, \alpha}$.

Przykład 3.

Jest to klasyczny przykład podany po raz pierwszy przez R.A. Fishera w 1936 r. i przytaczany odtąd przez wielu autorów. Niech y_1 będzie długością działki kielicha kwiatu, y_2 — szerokością działki kielicha, y_3 — długością płatka kwiatu i y_4 — szerokością płatka kwiatu. Mamy więc $p = 4$ zmienne.

Z populacji *Iris versicolor* wzięto $n_1 = 50$ obserwacji i tyleż $n_2 = 50$ obserwacji z populacji *Iris setosa*. Dane liczbowe (wyrażone w centymetrach) zestawiono w formie dwóch wektorów średnich

$$y_{1\cdot} = \begin{bmatrix} 5.936 \\ 2.770 \\ 4.260 \\ 1.326 \end{bmatrix}, \qquad y_{2\cdot} = \begin{bmatrix} 5.006 \\ 3.428 \\ 1.462 \\ 0.246 \end{bmatrix},$$

oraz estymatora macierzy kowariancji[5]

$$S = \begin{bmatrix} 0.195340 & 0.092200 & 0.099626 & 0.033055 \\ & 0.121079 & 0.047175 & 0.025251 \\ & & 0.125488 & 0.039586 \\ & & & 0.025106 \end{bmatrix}$$

Zgodnie ze wzorem (11.19) wyznaczamy wartość statystyki testowej

$$\tilde{F} = \frac{95}{4} \frac{50 \cdot 50}{100 \cdot 98} 103.2119 = 625.3256.$$

Wartość funkcji testowej jest większa od wartości granicznej $F_{0.01}$ odczytanej z tablic F Snedecora przy $\nu_1 = 4$ i $\nu_2 = 95$ stopniach swobody przy 1-procentowym poziomie istotności, więc odrzucamy hipotezę, że obie populacje mają jednakowe średnie wektory.

---

5 macierz jest symetryczna, wystarczy zatem podać jedynie jej górny trójkąt i główną przekątną

11.1.3 Wielowymiarowa miara dyskryminacyjna, funkcje dyskryminacyjne, dyskryminacja

Zdefiniujemy teraz miarę, która da wyraz temu, w jakim stopniu poziomy czynnika naruszają hipotezę zerową. Im większe są efekty działania czynnika, tym większe będzie naturalne zróżnicowanie w zachowaniu się struktur zmiennych. Dlatego miara wprowadzona niżej nazwana została miarą dyskryminacji, a dokładniej wielowymiarową miarą dyskryminacyjną p zmiennych. Miarą tą jest T^2, tzn.

$$T^2(y_1, ..., y_p) = tr(HG^{-1}).$$

W obecnie omawianym konkretnym przypadku dwóch zbiorowości otrzymujemy

$$T^2(y_1, ..., y_p) = \frac{1}{n_1 + n_2 - 2} \frac{n_1 n_2}{n_1 + n_2} (y_{1.} - y_{2.})^T S^{-1} (y_{1.} - y_{2.}) \tag{11.20}$$

Wielkość (11.20) określa, w jakim stopniu dane dwie próby przeczą hipotezie $\mu_1 = \mu_2$, innymi słowy jak duży jest wzajemny „odstęp statystyczny" omawianych dwóch populacji.

Rozważmy teraz określoną kombinację liniową danych p cech $y_1, ..., y_p$. Niech mianowicie

$$v = d_1 y_1 + d_2 y_2 + \ldots + d_p y_p, \tag{11.21}$$

gdzie wektor d o składowych $d_1, d_2, ..., d_p$ dany jest wzorem:

$$d = S^{-1} (y_{1.} - y_{2.}). \tag{11.22}$$

We wzorze (11.21) należy traktować wielkości $y_1, ..., y_p$, jak również v jako zmienne. Mamy więc

$$v = d^T y = (y_{1.} - y_{2.})^T S^{-1} y. \tag{11.23}$$

Cecha v jest nową cechą, którą na podstawie ostatnich równań można obliczyć dla każdego obiektu. Ma ona w obu rozważanych populacjach wartości średnie

$$v_1 = (y_{1.} - y_{2.})^T S^{-1} y_{1.}\ , \tag{11.24}$$

$$v_2 = (y_{1.} - y_{2.})^T S^{-1} y_{2.}\ , \tag{11.25}$$

a na rozrzut wewnątrzklasowy cechy v otrzymujemy

$$s_v^2 = (y_{1.} - y_{2.})^T S^{-1} (y_{1.} - y_{2.}). \tag{11.26}$$

Wielowymiarowa miara dyskryminacyjna cechy v jest równa

$$\begin{aligned} T^2 &= \frac{1}{n_1 + n_2 - 2} \frac{n_1 n_2}{n_1 + n_2} (v_{1.} - v_{2.})^2 \frac{1}{s_v^2} = \\ &= \frac{1}{n_1 + n_2 - 2} \frac{n_1 n_2}{n_1 + n_2} (y_{1.} - y_{2.})^T S^{-1} (y_{1.} - y_{2.}) \end{aligned} \tag{11.27}$$

tak że cecha v ma tę samą miarę dyskryminacyjną co wszystkie p cech pierwotnych razem wziętych. Spośród wszystkich kombinacji liniowych, które można utworzyć z p cech pierwotnych, największą miarę dyskryminacyjną ma cecha v.

Podana we wzorach (11.21) i (11.23) kombinacja liniowa cech $y_1, y_2, ..., y_p$ nazywa się funkcją dyskryminacyjną, a cechę v określamy jako cechę dyskryminacyjną. Wynik (11.27) orzeka, że rozróżnienie zbiorowości za pomocą cechy dyskryminacyjnej v jest tak samo możliwe, jak za pomocą p cech pierwotnych $y_1, y_2, ..., y_p$. Dlatego też przy dyskryminacji, tj. rozgraniczaniu (podziale) dowolnie danych obiektów na dwie klasy, będziemy stosowali jedną cechę dyskryminacyjną v zamiast p cech pierwotnych.

Przez dyskryminację (różnicowanie, diagnozowanie) można rozstrzygnąć, czy jakiś obiekt należy do klasy 1, czy do klasy 2. Zakłada się przy tym, że znamy wielkości n_1, n_2, y_1, y_2, S odpowiednio dla dwóch prób złożonych z rozważanych obiektów. W celach praktycznej realizacji różnicowania autorzy podają rozmaite reguły. Pewna trudność polega na tym, że różnicowanie nie może bazować na dokładnych parametrach rozkładu μ_1, μ_2, Σ, które są przecież nieznane lecz musi korzystać z odpowiednich ocen. W naszym postępowaniu wiążemy problem różnicowania (dyskryminacji) z pewnym problemem weryfikacyjnym. Mianowicie sprawdzamy dla $j = 1, 2$, czy dany obiekt z odpowiadającym mu wektorem y należy do zbiorowości j z wektorem wartości średnich $y_{j.}$, czy też nie. W tym celu oblicza się następujące wielkości testujące

$$k_1 = \frac{n_1}{n_1 + 1} (v - v_1)^2 \frac{1}{s_v^2}, \tag{11.28}$$

$$k_2 = \frac{n_2}{n_2 + 1} (v - v_{2.})^2 \frac{1}{s_v^2}, \tag{11.29}$$

gdzie wielkości v, $v_{1.}$, $v_{2.}$ i s_v^2 określone są wzorami (11.23) — (11.26). Przyjmujemy, że rozważany obiekt należy do zbiorowości j wtedy i tylko wtedy, gdy

$$k_j \leq F_{1, n_1+n_2-2, \alpha} = t^2_{n_1+n_2-2, \alpha} \tag{11.30}$$

gdzie α oznacza przyjęty poziom istotności.

Przy takim postępowaniu rozważany obiekt może być zaliczony do jednej z dwóch danych zbiorowości, do obu zbiorowości albo wreszcie do żadnej z nich. Za pomocą relacji (11.30) otrzymujemy wokół każdego z obu środków v_1 i v_2 pewien obszar rozrzutu (rozproszenia), który z prawdopodobieństwem $(1 - \alpha)$ zawiera w sobie obiekty rzeczywiście należące do rozpatrywanych klas.

Jeśli jednak chcemy w każdym przypadku podać jednoznaczne rozwiązanie problemu różnicowania, a więc jeśli szukamy najprawdopodobniejszej diagnozy, to trzeba się zdecydować na klasę o najmniejszej wartości k_j:

gdy $k_1 < k_2$, wtedy klasa 1,
gdy $k_1 > k_2$, wtedy klasa 2,
gdy $k_1 = k_2$, wtedy klasa 1 lub klasa 2.

W przypadku, gdy przy takiej dyskryminacji mają być uwzględnione tzw. prawdopodobieństwa aprioryczne p_1 i p_2, gdy zatem zakłada się z góry, z jakim prawdopodobieństwem dany obiekt należy do klasy 1 lub do klasy 2, wówczas dobrze jest korzystać z wielkości

$$l_j = \left(1 + \frac{k_j}{n_1 + n_2 - 2}\right)^{(n_1+n_2+1)/2} \cdot \frac{1}{p_j} \qquad (j = 1, 2) \tag{11.31}$$

Rozważany obiekt przydziela się do klasy z najmniejszą wartością l_j. W przypadku $p_1 = p_2$ ta reguła decyzyjna jest identyczna z uprzednią regułą.

Bardzo często pojawiającym się problemem jest możliwość zmniejszenia liczby rozpatrywanych cech pierwotnych[6]. Optimum jest bowiem uzyskanie przy możliwie najmniejszej liczbie cech możliwie największej miary dyskryminacyjnej.

W celu rozwiązania tego problemu określamy tzw. niezbędności poszczególnych cech y_i. Niezbędność U_i cechy y_i definiuje się jako wielkość, o jaką zmniejsza się miara dyskryminacyjna T^2, gdy ze zbioru wszystkich cech wyeliminuje się cechę y_i:

$$U_i = T^2(y_1, ..., y_p) - T^2(y_1, ..., y_{i-1}, y_{i+1}, ..., y_p) \qquad (i = 1, ..., p)$$

$$U_i = \frac{1}{n_1 + n_2 - 2} \frac{n_1 n_2}{n_1 + n_2} \frac{d_i^2}{t_{ii}}, \tag{11.32}$$

---

6 Proponujemy Czytelnikowi porównanie opisanego poniżej sposobu redukcji cech z metodami przedstawionymi w rozdziale dotyczącym regresji wielokrotnej. Wnioski mogą być pouczające!

gdzie t_{ii} oznacza i-ty element na głównej przekątnej macierzy $T = S^{-1}$. Jeśli chcemy po eliminacji pewnej cechy uzyskać możliwie największą miarę dyskryminacyjną, to należy eliminować tę cechę, której odpowiada najmniejsza wartość U_i.

Jeśli eliminowanych jest kolejno kilka cech, to po każdej kolejnej eliminacji należy niezbędności obliczać na nowo.

Istnieje pewien test istotności do weryfikacji hipotezy, że i-ta cecha jest cechą redundancyjną. Odpowiednia statystyka testowa ma postać:

$$\begin{aligned}\tilde{F}_i &= (n_1 + n_2 - p - 1)\,\frac{U_i}{1 + T^2(y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_p)} = \\ &= (n_1 + n_2 - p - 1)\,\frac{U_i}{1 + T^2(y_1, \ldots, y_p)}\end{aligned} \tag{11.33}$$

Hipotezę o redundancyjności odrzucamy jeżeli

$$\tilde{F} > F_{1,\, n_1 + n_2 - p - 1,\, \alpha}$$

Wielkość $\tilde{F}_i$ ma dokładnie rozkład F. Bezpośrednim wnioskiem z wzoru (11.33) jest to, że cecha o najmniejszej niezbędności U_i ma również najmniejszą wartość $\tilde{F}_i$.

Przykład 4.

Aby w porę zastosować odpowiednie postępowanie zapobiegawcze przy żółtaczce u noworodków, która polega na niedojrzałości wątroby, należy już w pierwszym dniu życia noworodka umieć przewidzieć, które dziecko może ewentualnie zachorować na żółtaczkę (np. w 5-tym dniu życia). W tym celu określono cztery wskaźniki liczbowe decydujące w różnym stopniu o możliwości zapadnięcia na żółtaczkę:

y_1 — wiek matki,
y_2 — waga noworodka,
y_3 — czas trwania ciąży,
y_4 — stopień zażółcenia.

Obiektami w naszym przykładzie są dzieci urodzone przedwcześnie, które rozdzielono na dwie klasy:

Klasa 1: Dzieci u których nie występuje żółtaczka noworodkowa i które z tego względu nie podlegają profilaktyce lekarskiej,

Klasa 2: Dzieci, które zapadły na tę chorobę, powiedzmy w 5-tym dniu życia i które trzeba poddać profilaktyce lekarskiej.

Zebrane dane dotyczące łącznie 31 nowo narodzonych dzieci są następujące:

$$n_1 = 20, \qquad n_2 = 11,$$
$$y_{1.} = [25{,}9 \quad 2184 \quad 266 \quad 1{,}78]^T,$$
$$y_{2.} = [23{,}5 \quad 2168 \quad 243 \quad 2{,}64]^T,$$

$$S = \begin{bmatrix} 26{,}4 & -63{,}7 & -26{,}6 & 0{,}362 \\ & 61973 & 2741 & -109 \\ & & 313 & -5{,}04 \\ & & & 0{,}527 \end{bmatrix}$$

Test do porównania wektorów wartości średnich obu klas noworodków daje

$$\tilde{F} = 9{,}05, \qquad \nu_1 = 4, \qquad \nu_2 = 26.$$

Przy poziomie istotności $\alpha = 0.05$ odczytujemy z tablic wartość $F_{4,26,0,05} = 2{,}74$. Zatem odnośnie do rozpatrywanych czterech cech istnieją istotne różnice między wcześniakami. Wyliczając wielowymiarową miarę dyskryminacyjną otrzymamy

$$T^2(y_1, y_2, y_3, y_4) = 1{,}39.$$

Funkcją dyskryminacyjną jest

$$v = 0{,}238v_1 - 0{,}0102v_2 + 0{,}139v_3 - 2.59v_4$$

Na wartości średnie wielkości v w obu klasach otrzymujemy

$$v_1 = 16{,}1, \qquad v_2 = 10{,}4,$$

tzn. że dzieci które zachorowały na żółtaczkę, mają średnio mniejsze wartości wielkości v aniżeli dzieci zdrowe. Aby przeprowadzić bezpośrednie sprawdzenie mocy dyskryminacyjnej cechy różnicującej v, można nasze 31 obiektów poddać postępowaniu dyskryminacyjnemu według (11.28) — (11.31) zarówno bez jak i wraz z uwzględnieniem prawdopodobieństw apriorycznych, które przyjmiemy odpowiednio do liczebności prób:

$$p_1 = 20/31 = 0{,}645, \qquad p_2 = 11/31 = 0{,}355.$$

Okazuje się, że rozróżnienie udaje się bardzo dobrze — w każdej klasie tylko jedno dziecko zostaje fałszywie zdiagnozowane. Różniące się wzajemnie prawdopodobieństwa aprioryczne nie wpływają na te wyniki.

Możemy również dokonać rozróżnienia posługując się tylko jedną cechą — czasem trwania ciąży (jest to cecha najmocniejsza). Wówczas w każdej klasie trzy obiekty będą

zdiagnozowane fałszywie, a przy uwzględnieniu prawdopodobieństw apriorycznych nawet jeszcze więcej. Różnicowanie jednowymiarowe jest więc wyraźnie gorsze od wielowymiarowego.

Obliczając następnie niezbędności wszystkich czterech cech otrzymamy w wyniku:

$$U_1 = 0{,}322, \quad U_2 = 0{,}697, \quad U_3 = 0{,}802, \quad U_4 = 0{,}543,$$

oraz na odpowiadające im wartości $\tilde{F}$:

$$\tilde{F}_1 = 4{,}04, \quad \tilde{F}_2 = 10{,}7, \quad \tilde{F}_3 = 13{,}1, \quad \tilde{F}_4 = 7{,}64.$$

Wartość krytyczna odczytana z tablic rozkładu F Snedecora wynosi tu $F_{1;26;0,05} = 4.23$. Okazuje się zatem, że cechy: waga noworodka (y_2), czas trwania ciąży (y_3) i stopień zażółcenia (y_4) nie są redundancyjne. Odnośnie do wieku matki (y_1) nie można odrzucić hipotezy o redundancyjności na poziomie $\alpha = 5\%$. Wypowiedź odnosząca się do wagi noworodka jest szczególnie interesująca, gdyż cecha ta z jednowymiarowego punktu widzenia jest prawie bezwartościowa.

Jeśli w kolejnych krokach redukcyjnych eliminowana będzie za każdym razem cecha o najmniejszej niezbędności, to najpierw znika y_1, potem y_4, a następnie y_2. Po eliminacji cechy y_1 mamy następujące niezbędności:

$$U_2 = 0{,}547, \quad U_3 = 0{,}561, \quad U_4 = 0{,}481,$$

po eliminacji cechy y_4 otrzymujemy niezbędności

$$U_2 = 0{,}206, \quad U_3 = 0{,}589,$$

a po eliminacji cechy y_2 otrzymujemy

$$U_3 = 0{,}384.$$

Analiza przykładu pokazuje zatem, że z odpadnięciem wieku matki również i czas trwania ciąży silnie traci na znaczeniu i że ze skreśleniem stopnia zażółcenia występuje znaczna strata w dyskryminacji odnośnie do wagi noworodka. Na takich właśnie zmianach niezbędności można rozpoznać wzajemne zależności między poszczególnymi cechami.

11.2 Wielowymiarowa analiza wariancji w przypadku wielu populacji i przy klasyfikacji pojedynczej

W poprzednim podrozdziale uzyskaliśmy ilościową relację dotyczącą odległości (odstępu) dwóch populacji statystycznych. Wielowymiarowa analiza wariancyjna i analiza dyskryminacyjna, które rozpatrywane są w tym podrozdziale, dają nam w przypadku więcej niż dwóch populacji nie tylko wzajemną odległość każdych dwóch zbiorowości, ale też pozwalają na zorientowanie się we wzajemnym położeniu wszystkich badanych zbiorowości. Obok testów istotności otrzymuje się rezultaty dotyczące struktury wielowymiarowych pomiarów.

11.2.1 Różnice wektorów wartości średnich

Podrozdział jest poświęcony badaniu różnic wartości średnich między wieloma populacjami. Rozpatrywać będziemy J populacji złożonych z poszczególnych obiektów. Zakładamy, że p-wymiarowe wektory wyników obserwacji, które należą do populacji j ($j = 1, 2, \ldots, J$) podlegają rozkładowi normalnemu $N(\mu_j, \Sigma)$, gdzie μ_j oznacza wektor wartości średnich klasy j, a Σ macierz kowariancji jednakową dla wszystkich populacji. Zarówno wektory μ_j, jak i macierz Σ są nieznane.

Zakładamy dalej, że z każdej populacji pobrano próbę złożoną z p-wymiarowych wektorów wyników obserwacji, przy czym próba taka odpowiadająca klasie j za każdym razem składa się z n_j wektorów wyników obserwacji. Poszczególne wektory wyników obserwacji klasy j oznaczamy jako:

$$y_{jk} = \begin{bmatrix} y_{1jk} \\ y_{2jk} \\ \cdot \\ \cdot \\ \cdot \\ y_{pjk} \end{bmatrix} \qquad (j = 1, \ldots, J;\ k = 1, \ldots, n_j)$$

Tabela danych pomiarowych będzie zatem miała postać jak poniżej.

Populacja 1: $y_{11} = [y_{111} \quad y_{211} \quad \ldots \quad y_{p11}]^T$

$y_{12} = [y_{112} \quad y_{212} \quad \ldots \quad y_{p12}]^T$

. . . . . .

$y_{1n_1} = [y_{11n_1} \quad y_{21n_1} \quad \ldots \quad y_{p1n_1}]^T$

Populacja 2: $y_{21} = [y_{121} \quad y_{221} \quad \ldots \quad y_{p21}]^T$
$y_{22} = [y_{122} \quad y_{222} \quad \ldots \quad y_{p22}]^T$
. . . . . .
$y_{2n_2} = [y_{12n_2} \quad y_{22n_2} \quad \ldots \quad y_{p2n_2}]^T$

Populacja 3: . . . . . .

Liczba populacji (prób) wynosi J ($J \geq 2$), przy czym muszą być spełnione warunki:

$$n_j \geq 1 \qquad (j = 1, 2, ..., J)$$
$$n = \sum_j n_j \geq p + J + 2$$

Podstawą wszystkich dalszych obliczeń są wektory wartości średnich

$$y_{j\cdot} = \frac{1}{n_j} \sum_{k=1}^{n_j} y_{jk} \qquad (j = 1, ..., J) \tag{11.34}$$

poszczególnych prób, wektor wartości średnich

$$y_{\cdot\cdot} = \frac{1}{n} \sum_j \sum_k y_{jk} = \frac{1}{n} \sum_k n_j y_{j\cdot} \tag{11.35}$$

wszystkich klas (średnia ogólna), oceny macierzy kowariancji

$$S_j = \frac{1}{n_j - 1} \sum_k (y_{jk} - y_{j\cdot})(y_{jk} - y_{j\cdot})^T \qquad (j = 1, ..., J) \tag{11.36}$$

poszczególnych prób i wreszcie uśredniona macierz kowariancyjna

$$S = \frac{1}{n - J} \sum_j (n_j - 1) S_j \ . \tag{11.37}$$

Wektory $y_{j\cdot}$ i macierz S są nieobciążonymi ocenami wielkości μ_j i Σ.
Jeśli oznaczymy symbolami s_{hi} elementy macierzy S, to wielkości

$$s_{ii} \quad \text{oraz} \quad r_{hi} = \frac{s_{hi}}{\sqrt{s_{hh}\, s_{ii}}}$$

są uśrednionymi wewnątrzgrupowymi rozrzutami i korelacjami. Obie macierze $S = [s_{hi}]$ i $R = [r_{hi}]$ są symetryczne, dzięki czemu wystarczy podanie wartości elementów powyżej, ewentualnie poniżej głównej przekątnej.

Przykład 1.

W przykładzie z nadczynnością gruczołów tarczycowych (Dodatek 4) otrzymuje się wektory wartości średnich:

$$y_{1.} = \begin{bmatrix} 89.3 \\ 90.6 \\ 83.8 \\ 70.7 \\ 1.90 \\ 31.1 \\ 37.2 \\ 43.9 \\ 41.0 \\ .246 \end{bmatrix}, \quad y_{2.} = \begin{bmatrix} 76.4 \\ 78.3 \\ 77.0 \\ 69.3 \\ 1.84 \\ 68.4 \\ 69.1 \\ 60.1 \\ 54.5 \\ 1.55 \end{bmatrix}, \quad y_{3.} = \begin{bmatrix} 85.3 \\ 87.7 \\ 73.6 \\ 68.2 \\ 3.01 \\ 29.2 \\ 38.5 \\ 50.1 \\ 48.2 \\ .107 \end{bmatrix}$$

oraz uśrednioną macierz kowariancji i macierz korelacji postaci:

$$S = \begin{bmatrix} 240 & 164 & 149 & 139 & 1.57 & 114 & 116 & 116 & 93.5 & 0.0877 \\ & 192 & 169 & 170 & -1.29 & 85.5 & 83.6 & 70.4 & 59.9 & -0.0571 \\ & & 192 & 207 & -3.79 & 79.9 & 90.3 & 91.1 & 78.5 & -0.2400 \\ & & & 907 & -5.87 & 79.6 & 82.8 & 96.5 & 77.7 & 0.1400 \\ & & & & 1.04 & 2.82 & 4.40 & 3.93 & 4.33 & 0.0401 \\ & & & & & 212 & 222 & 245 & 227 & -0.1080 \\ & & & & & & 262 & 281 & 269 & -0.2690 \\ & & & & & & & 372 & 332 & -0.2280 \\ & & & & & & & & 332 & 0.0985 \\ & & & & & & & & & 0.1220 \end{bmatrix}$$

$$R = \begin{bmatrix} 0.76 & & & & & & & & \\ 0.66 & 0.88 & & & & & & & \\ 0.51 & 0.70 & 0.85 & & & & & & \\ 0.10 & -.09 & -.27 & -.33 & & & & & \\ 0.51 & 0.42 & 0.40 & 0.31 & 0.19 & & & & \\ 0.46 & 0.37 & 0.40 & 0.29 & 0.27 & 0.94 & & & \\ 0.39 & 0.26 & 0.34 & 0.29 & 0.20 & 0.87 & 0.90 & & \\ 0.33 & 0.24 & 0.31 & 0.24 & 0.29 & 0.86 & 0.91 & 0.95 & \\ 0.02 & -.01 & -.05 & 0.02 & 0.11 & -.02 & -.05 & -.03 & 0.02 \end{bmatrix}$$

Model statystyczny omawianego niżej testu opisany jest równaniem

$$y_{jk} = \mu_j + \varepsilon_{jk} \qquad (j = 1, ..., J;\ k = 1, ..., n) \tag{11.38}$$

Będziemy weryfikowali hipotezę o równości wszystkich wektorów wartości średnich

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_J \ . \tag{11.39}$$

Sposób konstrukcji statystyki testowej jest bardzo zbliżony do metody opisywanej w poprzednim podrozdziale. Tworzymy mianowicie dwie macierze kwadratowe stopnia p:

$$H = \sum_{j=1}^{J} n_j (y_{j.} - y_{..}) (y_{j.} - y_{..})^T \tag{11.40}$$

oraz

$$G = \sum_{j=1}^{J} \sum_{k=1}^{n_j} (y_{jk} - y_{j.}) (y_{jk} - y_{j.})^T \ , \tag{11.41}$$

które są niezbędne dla testu wielowymiarowego.
Jako statystykę testową stosujemy wyrażenie

$$\tilde{F} = \frac{f_2 - p + 1}{f_1 p} tr\,(HG^{-1}) \quad \text{gdzie:} \quad f_1 = J - 1 \qquad f_2 = n - J$$

Statystyka ta, w przypadku prawdziwości hipotezy H_0, ma w przybliżeniu rozkład F. Ponieważ

$$S = \frac{1}{n - J} G \ , \tag{11.42}$$

więc na statystykę $\tilde{F}$ otrzymujemy ostatecznie wzór:

$$\begin{aligned} \tilde{F} &= \frac{n - J - p + 1}{(J - 1)\, p\, (n - J)} tr \left(\sum_{j=1}^{J} n_j (y_{j.} - y_{..}) (y_{j.} - y_{..})^T S^{-1} \right) = \\ &= \frac{n - J - p + 1}{(J - 1)\, p\, (n - J)} \sum_{j=1}^{J} n_j (y_{j.} - y_{..})^T S^{-1} (y_{j.} - y_{..}) \end{aligned} \tag{11.43}$$

Wzory przeliczeniowe na stopnie swobody są następujące:

$$\nu_1 = \begin{cases} \dfrac{(J-1)p\,(n-J-p)}{n-(J-1)p-2}, & \text{gdy } n-(J-1)\,p-2>0 \\ \infty & \text{gdy } n-(J-1)\,p-2\le 0 \end{cases} \tag{11.44}$$

$$\nu_2 = n - J - p + 1 \tag{11.45}$$

Hipotezę o równości J wektorów wartości średnich $\mu_1, \mu_2, \ldots, \mu_J$ odrzucamy, o ile

$$\tilde{F} > F_{\nu_1, \nu_2; \alpha}$$

gdzie $F_{\nu_1, \nu_2; \alpha}$ oznacza odczytaną z tablic rozkładu F Snedecora wartość krytyczną przy poziomie istotności α. Wyliczając według wzoru (11.44) wartość ν_1 możemy otrzymać liczbę ułamkową — należy wtedy zaokrąglić ν_1 albo przeprowadzić interpolację korzystając z tablic.

Przykład 2.

Kontunuujemy poprzedni przykład. W przypadku schorzenia tarczycy otrzymujemy

$$\tilde{F} = 5.30, \qquad \nu_1 = 200, \qquad \nu_2 = 11.$$

Odpowiadająca temu wartość F odczytana z tablic przy poziomie istotności $\alpha = 0{,}05$ wynosi $F_{200,11;0,05} = 2{,}43$. Wobec tego musimy odrzucić hipotezę o równości wszystkich trzech wektorów wartości średnich. ■

Test określony wzorem (11.43) pozwala na globalne oszacowanie różnic istniejących między wszystkimi J próbami, ale nie dostarcza on żadnej informacji, w jakim stopniu poszczególne populacje różnią się jedna od drugiej. Istnieje więc potrzeba wprowadzenia testów istotności, które dotyczą par populacji. Opierać się będziemy na wprowadzonym już (wzór (11.38)) modelu:

$$y_{jk} = \mu_j + \varepsilon_{jk} \qquad (j = 1, \ldots, J;\ k = 1, \ldots, n)$$

Jako hipotezę zerową rozpatrzymy

$$H_{l/m0} : \mu_l = \mu_m\ , \tag{11.46}$$

tzn. weryfikujemy, czy wektory wartości średnich populacji l oraz m są równe, czy też nie. Odpowiednią statystykę testową skonstruujemy natychmiast, jeśli wykorzystamy rozważania z poprzedniego podrozdziału (wzór (11.19)). Jedynym odchyleniem od opisywanego tam

sposobu jest to, że ocena macierzy Σ oparta jest teraz na J próbach i w konsekwencji obliczana jest nie ze wzorów (11.17) i (11.18), lecz na podstawie wzorów (11.36) i (11.37). A zatem jako statystykę testową otrzymujemy wyrażenie

$$\tilde{F}_{l/m} = \frac{n-J-p+1}{p\,(n-J)} \frac{n_l n_m}{n_l + n_m} (y_{l\cdot} - y_{m\cdot})^T S^{-1} (y_{l\cdot} - y_{m\cdot}) \tag{11.47}$$

ze stopniami swobody

$$\nu_1 = p, \qquad \nu_2 = n - J - p + 1. \tag{11.48}$$

Między klasami l oraz m zachodzi więc istotna różnica, jeżeli

$$\tilde{F}_{l/m} > F_{p,\,n-J-p+1;\,\alpha}\ . \tag{11.49}$$

Przykład 3.

W przypadku danych dotyczących nadczynności tarczycy otrzymujemy

$$\tilde{F}_{1/2} = 8{,}84, \quad \tilde{F}_{1/3} = 0{,}907, \quad \tilde{F}_{2/3} = 7{,}13, \quad \nu_1 = 10, \quad \nu_2 = 11\ .$$

Odczytana z tablic rozkładu F Snedecora wartość krytyczna wynosi $F_{10,11;0,05} = 2{,}85$. A zatem w pojedynczym porównaniu wyłania się istotna różnica między klasami 1 i 2, oraz istotna różnica między klasami 2 i 3. Nie uzyskujemy istotnej różnicy między klasami 1 i 3, co potwierdza nasze spostrzeżenia, gdyż są to klasy pacjentów, u których w początkowym okresie leczenie było pomyślne.

Na podstawie wzorów (11.47), (11.48) i (11.49) możemy dokonać porównań wektorów wartości średnich dla wszystkich par populacji. W ogólnym przypadku uzyskamy wówczas dla pewnej liczby par różnicę istotną, a dla pozostałych par różnicę nieistotną. Nasuwa się pytanie, na ile możliwe jest uogólnienie wyników otrzymanych drogą pojedynczych porównań dwóch klas? Identycznie jak w przypadku jednowymiarowym osąd kompleksowy na podstawie poszczególnych wyników nie jest prawidłowy. Innymi słowy, jeśli test pojedynczych porównań zastosowany trzykrotnie daje wyniki

$$\mu_1 \neq \mu_2, \quad \mu_1 \neq \mu_3, \quad \mu_2 \neq \mu_3,$$

to nie można na tej podstawie wnioskować, że wszystkie trzy wektory wartości średnich μ_1, μ_2, μ_3 są parami różne jeden od drugiego. Można to sformułować jeszcze inaczej: Jeśli powyższe trzy wyniki sprawdziliśmy każdy osobno przy jednakowym poziomie istotności, np. $\alpha = 0{,}05$, to nie możemy stąd wnioskować, że zweryfikowaliśmy

i wypowiedź globalną według której żadne dwa z tych trzech wektorów nie są równe przy $\alpha = 0{,}05$.

11.2.2 Wielowymiarowa miara dyskryminacyjna

Wielowymiarową miarę dyskryminacyjną cech $y_1, ..., y_p$ zdefiniowaliśmy już w poprzednim podrozdziale dla dwóch populacji. W przypadku J populacji określa się ją analogicznie:

$$T^2(y_1, ..., y_p) = tr(HG^{-1}) ,$$

lub też

$$T^2(y_1, ..., y_p) = \frac{1}{n-J} \sum_{j=1}^{J} n_j (y_{j.} - y_{..})^T S^{-1} (y_{j.} - y_{..}) \tag{11.50}$$

Wartość T^2 jest zawsze nieujemna ($T^2 \geq 0$), przy czym równość $T^2 = 0$ oznacza, że p danych cech zupełnie się nie nadaje do rozróżnienia rozważanych J populacji. Im większa jest wartość miary T^2, tym lepsze jest rozróżnienie tych J populacji za pomocą p cech.

Przykład 4.

Rozpatrujemy dalej schorzenie tarczycowe. Dokonując obliczeń zgodnie z wzorem (11.50) otrzymamy jako wielowymiarową miarę dyskryminacyjną wartość

$$T^2(y_1, ..., y_p) = 9{,}64 . \blacksquare$$

Bardzo istotną właściwością miary T^2 jest jej niezmienniczość względem dowolnych, liniowych i regularnych transformacji cech. Mianowicie, jeśli pierwotny wektor cech $y = [y_1, ..., y_p]^T$ zastąpimy wektorem cech $z = [z_1, ..., z_p]^T$ wyliczanym z niego przez liniową transformację regularną

$$z = U\, y,$$

to wtedy zachodzi równość

$$T^2(z_1, ..., z_p) = T^2(y_1, ..., y_p) .$$

Tak więc wielowymiarowa miara dyskryminacyjna może być traktowana jako coś, co charakteryzuje całą liniową przestrzeń cech i co nie jest czymś specyficznym dla różnych

p-wymiarowych kombinacji cech. Przykładowo, miara dyskryminacyjna zbioru złożonego z dwóch cech jest identyczna z miarą dyskryminacyjną zbioru w skład którego wchodzą dwie inne cechy, a mianowicie suma oraz różnica cech pierwotnych.

Wielowymiarową miarę dyskryminacyjną możemy obliczać dla pewnej części całej p-wymiarowej przestrzeni cech. Utwórzmy zatem nowy u-wymiarowy wektor cech z na podstawie starego p-wymiarowego wektora cech y. Formalnie odpowiada to wprowadzeniu transformacji liniowej opisanej macierzą U:

$$z_{(u,1)} = U^T_{(u,p)}\, y_{(p,1)}$$

Miarę dyskryminacyjną dla tej nowej przestrzeni cech wyliczamy korzystając z równości

$$z_{j.} = U^T y_{j.}\,, \qquad z.. = U^T y_{..}\,, \qquad S_z = U^T S U$$

i podstawiając je do wzoru (11.50):

$$\begin{aligned} T^2(z_1, ..., z_u) &= \frac{1}{n-J} \sum_{j=1}^{J} n_j (z_{j.} - z_{..})^T S_z^{-1} (z_{j.} - z_{..}) = \\ &= \frac{1}{n-J} \sum_{j=1}^{J} n_j (y_{j.} - y_{..})^T U (U^T S U)^{-1} U^T (y_{j.} - y_{..}) \end{aligned} \tag{11.51}$$

Nasuwa się tu spostrzeżenie, aby w ten sam sposób obliczać także miarę dyskryminacyjną każdego zbioru częściowego cech $y_1, ..., y_p$, a w szczególności miarę $T^2(y_i)$ poszczególnej cechy, jak też i miarę $T^2(y_h, y_i)$ dowolnej pary cech. Zachodzi przy tym równość:

$$T^2(y_i) = \frac{1}{(n-J)\, s_{ii}} \sum_j n_j (y_{ij.} - y_{i..})^2 = \frac{h_{ii}}{(n-J)\, s_{ii}} \tag{11.52}$$

gdzie h_{ii} i s_{ii} są i-tymi elementami głównej przekątnej macierzy H i macierzy

$$S = \frac{1}{n-J} G$$

Zwróćmy uwagę, że we wzorach (11.50) i (11.51) występują macierze odwrotne. Oznacza to, iż między rozważanymi cechami nie mogą zachodzić żadne zależności liniowe. Możemy ominąć ten warunek zakładając, że w przypadku istnienia takich zależności cechy redundancyjne będą najpierw eliminowane, a dopiero potem wyliczać będziemy wielowymiarową miarę dyskryminacyjną wg (11.50) lub (11.51). Uwzględniając te rozważania możemy napisać:

$$T^2(y_1, y_2) = T^2(y_1)\,, \qquad T^2(y_1, y_2, y_1 + y_2) = T^2(y_1, y_2)$$

Tak samo, jeżeli zwiększymy zbiór cech, to polepszy się wielowymiarowe rozróżnienie w sensie miary dyskryminacyjnej:

$$T^2(y_1) \leq T^2(y_1, y_2) \leq \ldots T^2(y_1, y_2, \ldots, y_p)$$

Przykład 5.

W kontynuowanym przez nas przykładzie wyliczamy macierz zawierającą miary dyskryminacyjne par cech (y_h, y_i). Jest ona następująca:

$$\{T^2(y_h, y_i)\} = \begin{array}{c|cccccccccc} & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ \hline & 0{,}112 & 0{,}138 & 0{,}191 & 0{,}139 & 0{,}276 & 2{,}08 & 1{,}25 & 0{,}374 & 0{,}309 & 2{,}59 & 1 \\ & & 0{,}128 & 0{,}411 & 0{,}229 & 0{,}291 & 1{,}87 & 1{,}13 & 0{,}330 & 0{,}290 & 2{,}52 & 2 \\ & & & 0{,}093 & 0{,}249 & 0{,}222 & 1{,}55 & 0{,}99 & 0{,}297 & 0{,}262 & 2{,}48 & 3 \\ & & & & 0{,}003 & 0{,}170 & 1{,}24 & 0{,}71 & 0{,}138 & 0{,}112 & 2{,}41 & 4 \\ & & & & & 0{,}160 & 1{,}35 & 0{,}89 & 0{,}282 & 0{,}256 & 2{,}64 & 5 \\ & & & & & & 1{,}10 & 1{,}46 & 2{,}70 & 2{,}66 & 3{,}58 & 6 \\ & & & & & & & 0{,}63 & 1{,}50 & 1{,}94 & 3{,}16 & 7 \\ & & & & & & & & 0{,}116 & 0{,}133 & 2{,}56 & 8 \\ & & & & & & & & & 0{,}095 & 2{,}49 & 9 \\ & & & & & & & & & & 2{,}41 & \end{array}$$

Analiza danych zawartych w macierzy prowadzi do dwóch wniosków:

po pierwsze — najlepiej dyskryminującą cechą jest cecha y_{10}, natomiast najlepiej dyskryminującą parą cech jest (y_{10}, y_6),

po drugie — miara dyskryminacyjna w niektórych przypadkach tylko bardzo niewiele się zwiększa przy połączeniu dwóch cech pojedynczych w jedną cechę, (np. y_1 i y_2), a w innych przypadkach wzrasta w sposób znaczący (np. dla y_3 i y_4, dla y_4 i y_6). Korzystając z wzorów (11.43) — (11.45) możemy sprawdzić istotność miary T^2. Mianowicie dla $p = 1$ (tzn. pojedynczej cechy, czego odpowiednikiem są elementy leżące na głównej przekątnej) mamy $\tilde{F} = 10T^2$, $\nu_1 = 2$, $\nu_2 = 20$, a wartość krytyczna F wynosi $F = 0{,}49$ przy $\alpha = 0{,}5$. Natomiast jeśli $p = 2$ (tzn. dla par cech) otrzymujemy $\tilde{F} = \frac{19}{4} T^2$, $\nu_1 = \frac{79}{17}$, $\nu_2 = 19$, oraz wartość krytyczna dla $\alpha = 0{,}05$ wynosi w przybliżeniu $F = 0{,}602$. ■

Rozpatrzmy teraz dwa zbiory cech: zbiór $\mathbf{Z}_1$ zawierający p_1 cech oraz $\mathbf{Z}_2$ zawierający p_2 cech. Załóżmy dalej, że jesteśmy w stanie określić wielowymiarową miarę dyskryminacyjną zarówno dla zbioru $\mathbf{Z}_1$, jak i dla $\mathbf{Z}_2$ oraz dla sumy tych zbiorów $\mathbf{Z}_1 \cup \mathbf{Z}_2$. Może wówczas zajść:

1. $T^2(\mathbf{Z}_1 \cup \mathbf{Z}_2) = T^2(\mathbf{Z}_1)$
Miara dyskryminacyjna nie powiększa się przez dołączenie zbioru cech $\mathbf{Z}_2$ do zbioru cech $\mathbf{Z}_1$, czyli zbiór cech $\mathbf{Z}_2$ jest redundancyjny w stosunku do zbioru cech $\mathbf{Z}_1$.

2. $T^2(\mathbf{Z}_1 \cup \mathbf{Z}_2) = T^2(\mathbf{Z}_1) = T^2(\mathbf{Z}_2)$
Oba zbiory są redundancyjne jeden w stosunku do drugiego, tzn. są równoważne w swych możliwościach diagnostycznych i każdy z nich może być całkowicie zastąpiony przez drugi.

3. $T^2(\mathbf{Z}_1 \cup \mathbf{Z}_2) = T^2(\mathbf{Z}_1) + T^2(\mathbf{Z}_2)$
Każdy ze zbiorów wnosi do dyskryminacji populacji swój wkład niezależnie od wkładu drugiego zbioru.

4. $T^2(\mathbf{Z}_1 \cup \mathbf{Z}_2) > T^2(\mathbf{Z}_1) = T^2(\mathbf{Z}_2)$
Przez połączenie obu zbiorów $\mathbf{Z}_1$ i $\mathbf{Z}_2$ otrzymujemy nadzwyczajne powiększenie wielowymiarowej miary dyskryminacyjnej. Nawet, gdy oba te zbiory cech, traktowane osobno, mają względnie niską miarę dyskryminacyjną, to jednak ich kombinacja może mieć dużą moc diagnostyczną. Właśnie dopiero kombinacja zbiorów cech ma swe pełne znaczenie, natomiast poszczególne te zbiory są stosunkowo mało znaczące.

11.2.3 Cechy dyskryminacyjne i funkcje dyskryminacyjne

Wyliczone na podstawie zebranych danych cechy dyskryminacyjne mogą zostać użyte do rozróżnienia (dyskryminacji) dowolnych obiektów, tzn. do podziału obiektów na grupy lub do diagnozowania obiektów w wydzielonych J grupach. Możliwe jest jednak także inne zastosowanie cech dyskryminacyjnych — za ich pomocą przenosimy zależności wynikłe w eksperymencie z pierwotnej p-wymiarowej przestrzeni do przestrzeni o mniejszej liczbie wymiarów.

Wiadomo, że wraz z powiększaniem przestrzeni cech także i odpowiednia miara dyskryminacyjna osiąga wartości większe (lub przynajmniej nie mniejsze) i że przy przejściu od danej przestrzeni cech do którejś z jej podprzestrzeni miara dyskryminacyjna na ogół zmniejsza się. Jednakże przestrzeń cech generowana przez cechy dyskryminacyjne mimo występującego zmniejszenia wymiaru zachowuje wielowymiarową miarę dyskryminacyjną pierwotnej przestrzeni.

Analogicznie jak w poprzednim podrozdziale określamy cechy dyskryminacyjne jako

$$v_j = (y_{j\cdot} - y_{\cdot\cdot})\, S^{-1}\, y \qquad (j = 1, \ldots, J) \tag{11.53}$$

gdzie wektor y oznacza zmienny, całkowicie nieokreślony p-wymiarowy wektor wartości pomiarowych, natomiast $y_{j\cdot}$, $y_{\cdot\cdot}$ oraz S określone zostały przez dane J prób. Cechy v_j

noszą nazwę elementarnych cech dyskryminacyjnych, a określone wzorem (11.53) przepisy obliczeniowe do uzyskania wielkości v_j z wielkości y_i nazywają się elementarnymi funkcjami obliczeniowymi. Określenie *elementarne* oznacza w tym przypadku, że cechy v_j mogą być obliczane za pomocą prostych operacji macierzowych, bez konieczności rozwiązywania jakiegokolwiek zagadnienia własnego.
Przyjmując

$$d_j = S^{-1}(y_{j\cdot} - y_{\cdot\cdot}) \ ,$$

$$A_{(p,J)} = (y_{1\cdot} - y_{\cdot\cdot}, y_{2\cdot} - y_{\cdot\cdot}, \ldots, y_{J\cdot} - y_{\cdot\cdot}) \tag{11.54}$$

oraz

$$D_{(p,J)} = [d_1, d_2, \ldots, d_J] = S^{-1}A \ , \tag{11.55}$$

mamy dla cechy v_j

$$v_j = d_j^T y = \sum_{i=1}^{p} d_{ij} y_i \tag{11.56}$$

i dla wektora v wszystkich elementarnych cech dyskryminacyjnych otrzymuje się

$$v = D^T y \ . \tag{11.57}$$

Elementarne cechy dyskryminacyjne nie są wszystkie wzajemnie liniowo niezależne, zachodzi bowiem relacja

$$\sum_j n_j v_j = 0 \ . \tag{11.58}$$

Dowodzi się także, że wielowymiarowa miara dyskryminacyjna elementarnych cech dyskryminacyjnych jest równa mierze dyskryminacyjnej pierwotnych cech:

$$T^2(v_1, v_2, \ldots, v_J) = T^2(y_1, y_2, \ldots, y_p) \ . \tag{11.59}$$

Interpretacja powyższego stwierdzenia jest następująca: elementarne cechy dyskryminacyjne odnośnie podziału na grupy i przeprowadzenia dyskryminacji są tak samo cenne jak p cech pierwotnych. Jeżeli liczba populacji J nie przekracza p, to wtedy istnieje $J - 1$ liniowo niezależnych elementarnych cech dyskryminacyjnych — zatem dzięki transformacji (11.57) liczba rozważanych cech zmniejsza się.

Poszczególne elementarne cechy dyskryminacyjne również same posiadają pewną interpretację. Mianowicie cecha v_j nadaje się szczególnie dobrze do tego, aby wyodrębnić populację j z całego zbioru wszystkich populacji.

Przykład 6.

W rozważanym przypadku ze schorzeniem tarczycowym otrzymujemy następujące elementarne funkcje dyskryminacyjne:

$$v_1 = 0{,}0324y_1 + 0{,}0387y_2 - 0{,}00114y_3 - 0{,}0164y_4 - 0{,}0162y_5 - 0{,}0866y_6 - \\ - 0{,}0669y_7 + 0{,}0124y_8 + 0{,}0799y_9 - 1{,}96y_{10},$$

$$v_2 = -0{,}136y_1 - 0{,}4360y_2 + 0{,}383y_3 - 0{,}0489y_4 - 0{,}268y_5 + 0{,}635y_6 + \\ + 0{,}252y_7 - 0{,}117y_8 - 0{,}452y_9 + 11{,}0y_{10},$$

$$v_3 = 0{,}00906y_1 + 0{,}376y_2 - 0{,}504y_3 + 0{,}153y_4 + 0{,}444y_5 - 0{,}385y_6 + \\ + 0{,}0208y_7 + 0{,}0898y_8 + 0{,}177y_9 - 4{,}26y_{10}. \blacksquare$$

Oprócz elementarnych cech dyskryminacyjnych istnieją również tzw. nieelementarne cechy dyskryminacyjne. Otrzymuje się je poprzez rozwiązanie zagadnienia własnego

$$\frac{1}{n-J} He = \lambda Se \tag{11.60}$$

Jeśli $\lambda_1, \lambda_2, \ldots \lambda_t$ ($\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_t$) są różnymi od zera i wzajemnie różnymi wartościami własnymi zagadnienia własnego (11.60), a $e_1, e_2, \ldots, e_t$ odpowiadającymi im unormowanymi wektorami własnymi, to wówczas cechy

$$w_h = e_h^T y = \sum_{i=1}^{p} e_{ih}\, y_i \qquad (h = 1, \ldots, t) \tag{11.61}$$

nazywają się nieelementarnymi cechami dyskryminacyjnymi. Liczba t oznacza tu rząd macierzy H. Zachodzi

$$t = \min(p, J - 1) \tag{11.62}$$

tzn., że liczba nieelementarnych cech dyskryminacyjnych nie przekracza liczby cech pierwotnych i jest co najmniej o jeden mniejsza niż liczba populacji. Zgodnie z tym przy dwóch klasach mamy tylko jedną nieelementarną cechę dyskryminacyjną. Przyjmując

$$E_{(p,t)} = [e_1, e_2, \ldots, e_t] \tag{11.63}$$

otrzymujemy w konsekwencji dla wektora w wszystkich nieelementarnych cech dyskryminacyjnych

$$w = E^T y \ . \tag{11.64}$$

Liczby e_{ih} są współczynnikami wagowymi nieelementarnych funkcji dyskryminacyjnych. Dla macierzy E zachodzi warunek ortonormalności:

$$E^T SE = I \ . \tag{11.65}$$

Cechy w_h są wzajemnie liniowo niezależne i rozpinają one tę samą przestrzeń cech co i elementarne cechy dyskryminacyjne (dowód tego twierdzenia pomijamy). Zgodnie ze wzorem (11.59) ich miara dyskryminacyjna jest więc równa mierze dyskryminacyjnej p cech pierwotnych:

$$T^2 (w_1, w_2, ..., w_t) = T^2 (y_1, y_2, ..., y_p) \ . \tag{11.66}$$

Nieelementarne cechy dyskryminacyjne, oprócz tego, że dają co najmniej tyle samo co i elementarne cechy dyskryminacyjne, posiadają również pewne własności optymalizacyjne. Mianowicie cecha w_1 ma spośród wszystkich cech, które mogą powstać z y_i przez tworzenie kombinacji liniowych, maksymalną miarę dyskryminacyjną. Spośród wszystkich par cech, które można utworzyć za pomocą transformacji liniowych pierwotnych cech y_i, para w_1, w_2 ma największą miarę dyskryminacyjną, itd. Cecha w_1 jest cechą najlepiej dyskryminacyjną, cecha w_2 jest najlepszym uzupełnieniem do w_1 , cecha w_3 jest najlepszym uzupełnieniem do pary cech w_1, w_2, ... itd. Dlatego też bardzo często przedstawia się wizualnie wyniki dyskryminacji korzystając z płaszczyzny utworzonej przez cechy w_1 oraz w_2.

Dla $h = 1, \ldots , t$ zachodzi

$$T^2 (w_h) = \lambda_h \ , \tag{11.67}$$

$$T^2 (w_1, w_2, ..., w_h) = \lambda_1 + \lambda_2 + \ldots + \lambda_h \ , \tag{11.68}$$

a więc wartości własne λ_h należy interpretować jako miary dyskryminacyjne odpowiednich cech dyskryminacyjnych. Wielowymiarowa miara dyskryminacyjna wielu nieelementarnych cech dyskryminacyjnych równa jest sumie odpowiadających im wartości własnych.

Przykład 7.

W dalszym ciągu rozpatrujemy nasz poprzedni przykład. Otrzymujemy

$$t = \min(10, 2) = 2$$

oraz

$$\lambda_1 = 9{,}09\,, \qquad \lambda_2 = 0{,}547\,.$$

Natomiast wyliczone nieelementarne cechy dyskryminacyjne są następujące:

$$w_1 = -0{,}0214y_1 - 0{,}0772y_2 + 0{,}0721y_3 - 0{,}0115y_4 - 0{,}0528y_5 + 0{,}108y_6 + \\ + 0{,}0386y_7 - 0{,}0204y_8 - 0{,}0746y_9 + 1{,}82y_{10}\,,$$

$$w_2 = -0{,}0274y_1 + 0{,}107y_2 - 0{,}192y_3 + 0{,}0742y_4 + 0{,}185y_5 - 0{,}065y_6 + \\ + 0{,}0717y_7 + 0{,}0225y_8 - 0{,}00855y_9 + 0{,}234y_{10}\,.$$

Nasunąć się może pytanie: ile nieelementarnych cech dyskryminacyjnych należy uważać za statystycznie istotne? Istnieje pewien test pozwalający na sprawdzenie tej hipotezy przy warunku posiadania danych pomiarowych o dużej liczebności (duża wartość n). Jeśli mianowicie mamy sprawdzić, czy ostatnie $t - t_1$ cech dyskryminacyjnych w_{t_1+1}, w_{t_1+2}, ..., w_t mają nieistotnie odchylającą się od zera miarę dyskryminacyjną, to należy wziąć pod uwagę wyrażenie

$$\chi^2 = (n - J - p + t_1 + 1)\,(\lambda_{t_1+1} + \lambda_{t_1+2} + \ldots + \lambda_t)\,. \tag{11.69}$$

W przypadku, gdy wartość χ^2 nie jest większa niż odczytana z tablic wartość krytyczna rozkładu χ^2 dla odpowiedniego poziomu α i przy $(J - t_1 - 1)(p - t_1)$ stopniach swobody, wówczas wnioskujemy, że te ostatnie $t - t_1$ cechy dyskryminacyjne powinniśmy uznać za statystycznie nieistotne.

Aby dla konkretnych danych uzyskać liczbę statystycznie istotnych wymiarów przestrzeni dyskryminacyjnej, należy ten test przeprowadzić kolejno dla $t_1 = 0$, $t_1 = 1$, ..., i wreszcie $t_1 = t - 1$.

Gdy dla $t_1 < t^*$ otrzymamy istotne odchylenie od zera, natomiast dla $t_1 \geq t^*$ — już nie, wówczas to t^* daje nam szukany wymiar przestrzeni dyskryminacyjnej.

Przykład 8.

Dla danych dotyczących schorzenia tarczycowego obliczamy wartość χ^2 dla drugiej wartości własnej $\lambda_2 = 0{,}547$ (patrz poprzedni przykład). Otrzymujemy

$$\chi^2 = 12 \cdot 0{,}547 = 6{,}56 \ .$$

Odczytana z tablic wartość krytyczna $\chi^2_{9;0,05} = 16{,}9$. A zatem druga cecha dyskryminacyjna nie różni się w sposób istotny od zera.

11.2.4 Przeprowadzanie dyskryminacji

Zaczniemy od przypadku, kiedy do dyskryminacji wykorzystujemy wszystkie t nieelementarnych cech dyskryminacyjnych. Zakładamy, że dla określonego obiektu mamy już obliczone t wartości w_1, w_2, ..., w_t, tzn. wektor w według (11.61) lub (11.64). Dyskryminacja opierać się będzie na odpowiednim teście istotności, który daje możliwość sprawdzenia, czy wyliczony przez nas wektor w jest reprezentantem pewnej dalszej populacji, której możemy nadać numer 0, obok istniejących już populacji 1, 2, ..., J.Biorąc za przykład wzór (11.47) otrzymujemy jako wyrażenie testowe

$$\begin{aligned} \tilde{F}_{0/j} = k_j &= \frac{n-J-t+1}{t\,(n-J)} \ \frac{n_j}{n_j+1}\,(w-w_{j\cdot})^T\,(w-w_{j\cdot}) = \\ &= \frac{n-J-t+1}{t\,(n-J)} \ \frac{n_j}{n_j+1} \sum_{h=1}^{t} (w_h - w_{hj\cdot})^2 \end{aligned} \tag{11.70}$$

W powyższym wzorze nie występuje jawnie macierz kowariancyjna, a to dlatego, że nieelementarne cechy dyskryminacyjne mają macierz jednostkową jako macierz kowariancyjną[7]. Symbolem $w_{j\cdot}$ oznaczamy wektor wartości średnich klasy j, a obliczamy go ze wzoru

$$w_{j\cdot} = E^T y_{j\cdot} \ , \tag{11.71}$$

przy czym $w_{hj\cdot}$ oznaczają jego poszczególne składowe. Identycznie jak w przypadku poprzednich testów będziemy uważać, że wektor w (a co za tym idzie przyporządkowany mu obiekt) należy do j-tej klasy, jeżeli

$$k_j \le F_{t,\,n-J-t+1;\,\alpha} \tag{11.72}$$

Wielkość $F_{t,\,n-J-t+1;\,\alpha}$ oznacza oczywiście odpowiednią wartość krytyczną, odczytaną z tablic rozkładu F przy założonym z góry prawdopodobieństwie popełnienia błędu wynoszącym α.

---

7 porównaj zależność (11.65) !

Może się zdarzyć, że pewien obiekt zostanie przyporządkowany jednocześnie wielu populacjom, albo że nie zostanie w ogóle przyporządkowany. Nierówność (11.72) oznacza, że każdej populacji przyporządkowany zostaje t-wymiarowy kulisty obszar rozrzutu, który z prawdopodobieństwem $1 - \alpha$ zawiera w sobie obiekty istotnie należące[8] do danej populacji. Możemy zlikwidować tę wieloznaczną dyskryminację przyporządkowywując każdy obiekt dokładnie jednej populacji diagnozując na podstawie każdorazowego największego prawdopodobieństwa, tzn. wybierając populację o najmniejszej wartości k_j.

Jeżeli interesuje nas najlepsza diagnoza przy warunku ubocznym, że każdej populacji przypisane jest z góry pewne prawdopodobieństwo p_j (prawdopodobieństwo *a priori*), tzn. gdy już z góry ma być uwzględnione, z jakim prawdopodobieństwem dany obiekt trafia do odpowiedniej populacji, to do takiej dyskryminacji stosuje się wielkości

$$l_j = \left(1 + \frac{1}{n - J - t + 1}\right)^{(n+1)/2} \cdot \frac{1}{p_j} \tag{11.73}$$

przy czym dany obiekt zostaje przyporządkowany populacji o najmniejszej wartości l_j. Jeśli wszystkie prawdopodobieństwa p_j będą jednakowe, to dyskryminacja ta jest identyczna z poprzednią opierającą się na wartościach k_j.

Przykład 9.

Przeprowadźmy dyskryminację dla naszych 23 obiektów zgrupowanych w trzy klasy. Prawdopodobieństwa aprioryczne są wybrane odpowiednio do wielkości rozważanych prób, tzn.:

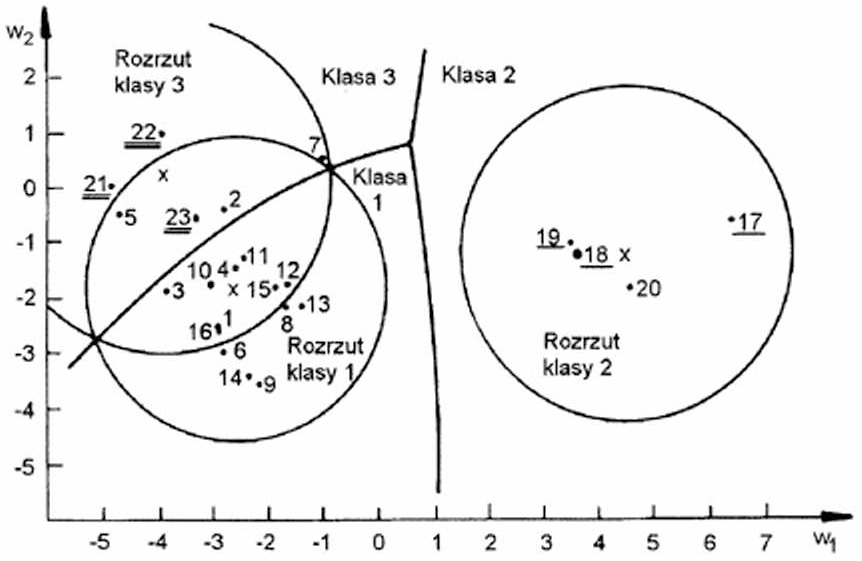
$$p_1 = \frac{16}{23} = 0{,}696\,, \qquad p_2 = \frac{4}{23} = 0{,}174\,, \qquad p_3 = \frac{3}{23} = 0{,}130.$$

Graficznie wyniki dyskryminacji prezentuje rysunek 11.2. Widoczne koła przedstawiają obszary rozrzutu poszczególnych klas (według (11.72)), natomiast linie graniczne między klasami umożliwiają dyskryminację z możliwie największym prawdopodobieństwem (tzn. dla każdego obiektu można odczytać klasę o najmniejszym k_j). Zaznaczone punkty odpowiadają poszczególnym obiektom, natomiast podkreślenie numerów ma na celu wskazanie podziału na klasy. ■

Dotychczas w procesie dyskryminacji opieraliśmy się na nieelementarnych cechach dyskryminacyjnych $w_1, \ldots, w_t$. Takie same wyniki można również uzyskać za pomocą

---

8 tzn. dla $\alpha = 0{,}05$ w obszarze leży 95% obiektów



Rys. 11.2 Dyskryminacja obiektów w schorzeniu tarczycowym.

elementarnych cech dyskryminacyjnych $v_1, \ldots, v_J$. Ponieważ nie wszystkie cechy v_j są nawzajem liniowo niezależne, więc do dyskryminacji bierzemy tylko wektor $v^\star$ pierwszych t cech dyskryminacyjnych

$$v^\star = (v_1, \ldots, v_t)^T$$

W przypadku gdy $J \leq p + 1$, mamy $t = J - 1$ i wtedy wykorzystuje się wszystkie elementarne cechy dyskryminacyjne, za wyjątkiem ostatniej.

Zamiast wzoru (11.70) otrzymujemy teraz następującą równość

$$k_j = \frac{n - J - t + 1}{t\,(n - J)} \cdot \frac{n_j}{n_j + 1} (v^\star - v^\star_{j\cdot})^T S^{-1}_{v^\star} (v^\star - v^\star_{j\cdot}) \tag{11.74}$$

Macierz $S^{-1}_{v^\star} = A^{\star T} S^{-1} A^\star$ jest macierzą kowariancyjną wektora cech $v^\star$, macierz $A^\star$ jest macierzą zbudowaną z pierwszych t kolumn macierzy A, natomiast dla wektorów wartości średnich zachodzi relacja:

$$v^\star_{j\cdot} = A^{\star T} S^{-1} y_{j\cdot} \tag{11.75}$$

O równoważności wzorów (11.70) i (11.74) świadczy poniższa tożsamość (wyprowadzenie pomijamy):

$$(w - w_{j.})^T (w - w_{j.}) = (v^{\star} - v_{j.}^{\star})^T S_{v^{\star}}^{-1} (v^{\star} - v_{j.}^{\star}) \tag{11.76}$$

Zależność (11.72) pozwalała nam na przyporządkowywanie obiektów do poszczególnych populacji. Dążyliśmy przy tym do tego, aby w miarę możliwości nie uzyskiwać żadnych wieloznacznych wypowiedzi. Wieloznaczności można uniknąć, gdy obszary rozrzutu odpowiadające poszczególnym klasom wzajemnie się nie przecinają. Możemy w stosunkowo prosty sposób, nawiązując do problemu porównywania wektorów wartości pomiarowych, sprawdzić, czy obszary rozrzutu odpowiadające dwóm danym populacjom l oraz m mają, czy też nie mają wspólnych punktów.
Obszary te nie przecinają się wzajemnie jedynie wtedy, gdy punkt krytyczny

$$\overline{w}_{l/m} = \frac{r_l w_{l.} + r_m w_{m.}}{r_l + r_m} ,$$

który jest punktem środkowym między $w_{l.}$ a $w_{m.}$, gdzie

$$r_l = \sqrt{\frac{n_l + 1}{n_l}} , \qquad r_m = \sqrt{\frac{n_m + 1}{n_m}}$$

są promieniami tych kulistych obszarów rozrzutu, leży poza obiema kulami. Oznacza to, że musi zachodzić nierówność:

$$k_{l/m} = \frac{n - J - t + 1}{t\,(n - J)} \frac{1}{(r_l + r_m)^2} (w_{l.} - w_{m.})^T (w_{l.} - w_{m.}) > F_{t,\, n-J-t+1;\, \alpha}$$

Nierówność tę możemy zapisać w postaci niewymagającej użycia wektorów cech dyskryminacyjnych:

$$k_{l/m} = \frac{n - J - t + 1}{t\,(n - J)} \frac{1}{(r_l + r_m)^2} (y_{l.} - y_{m.})^T S^{-1} (y_{l.} - y_{m.}) > F_{t,\, n-J-t+1;\, \alpha} \tag{11.77}$$

Dwie populacje, dla których zachodzi warunek (11.77) będziemy uważali za wzajemnie odizolowane. Korzyść z obliczania wielkości $k_{l/m}$ polega na tym, że bez uprzedniego przeprowadzenia analizy dyskryminacyjnej można już uzyskać pewne informacje o ewentualnych oczekiwanych wieloznacznościach.
Warto podkreślić, że wprowadziliśmy już dwie metody rozróżnialności dwóch populacji:
1. Porównania pojedyncze według (11.47) — (11.49),
2. Weryfikacja izolowalności według (11.77).

Przykład 10.

W przypadku schorzenia tarczycowego otrzymujemy

$$k_{1/2} = 5{,}17\ , \qquad k_{1/3} = 0{,}649\ , \qquad k_{2/3} = 6{,}96.$$

Odpowiadająca temu wartość krytyczna odczytana z tablic rozkładu F wynosi $F_{2,19;0,05} = 3{,}52$. Populacje 1 i 2, jak również populacje 2 i 3 są wzajemnie odizolowane. Obiekt, który w procesie dyskryminacji został przydzielony do populacji 2, nie może więc już być przypisany do żadnej z pozostałych dwóch populacji. ■

Jeśli do dyskryminacji ma być wykorzystanych mniej niż t cech dyskryminacyjnych, przykładowo t^*, to wtedy zgodnie z uprzednimi rozważaniami wykorzystujemy nieelementarne cechy dyskryminacyjne $w_1, w_2, \ldots, w_{t^*}$. W takim przypadku można również stosować wzory (11.70) i (11.72) zastępując t wartością t^*. Na ogół sensownie jest nie uwzględniać słabo rozgraniczających cech dyskryminacyjnych. Uwzględnianie nieefektywnych cech dyskryminacyjnych niepotrzebnie zwiększa nakład pracy obliczeniowej w analizie dyskryminacyjnej i powoduje niekiedy nawet wzrost błędów w dyskryminacji.

Przy eliminacji mało informatywnych cech dyskryminacyjnych najlepiej jest oprzeć się na teście wymiaru zgodnie z (11.69). Eliminowane są wtedy wszystkie cechy dyskryminacyjne statystycznie nieistotne.

Dyskryminacja prowadzona z mniejszą niż t liczbą elementarnych cech dyskryminacyjnych może mieć sens tylko w niektórych specyficznych przypadkach. Zaznaczmy jeszcze, że wzór (11.77) na wzajemną izolowalność dwóch populacji nie stosuje się do dyskryminacji z mniejszą niż t liczbą cech dyskryminacyjnych.

11.2.5 Eliminacja zbędnych cech

Planując konkretne badanie staramy się z reguły zebrać jak najwięcej informacji dla późniejszej analizy statystycznej. Oznacza to, że dążymy do wprowadzenia i pomierzenia jak największej liczby cech. W obliczeniach z kolei duże znaczenie ma określenie wysoce efektywnych kombinacji cech. Dąży się przy tym do tego, aby uzyskać przy możliwie niewielkiej liczbie cech możliwie dużą miarę dyskryminacyjną. A zatem w procesie analizy staramy się z kolei zredukować liczbę występujących cech, zostawiając jedynie te najbardziej przydatne (oczywiście z punktu widzenia przyjętego kryterium).

Redukcję cech można przeprowadzać stopniowo. Zmniejsza się przy tym liczbę cech kolejno z p do $(p-1)$, do $(p-2)$, itd. W każdym kolejnym kroku eliminowana jest ta cecha, która przez swe wyłączenia powoduje najmniejsze zmniejszenie wielowymiarowej miary dyskryminacyjnej. Jeśli wyjdziemy od p cech, to wówczas wyeliminowana zostanie cecha o najmniejszej niezbędności

$$U_i = T^2(y_1, ..., y_p) - T^2(y_1, ..., y_{i-1}, y_{i+1}, ..., y_p) \tag{11.78}$$

Obliczanie poszczególnych niezbędności prowadzi się zgodnie z poniższym wzorem

$$U_i = \frac{1}{(n-J)\, t_{ii}} \sum_{j=1}^{J} n_j d_{ij}^2 , \tag{11.79}$$

gdzie d_{ij} są współczynnikami wagowymi elementarnych funkcji dyskryminacyjnych, a t_{ii} oznacza i-ty element diagonalny macierzy

$$T = S^{-1} .$$

W podobny sposób można postępować także w kolejnych następnych krokach redukcji. W każdym kolejnym kroku owe niezbędności U_i należy obliczać od nowa.

Redukcja cech może też być sterowana przez odpowiednią statystykę testową $\tilde{F}_i$, która odpowiada testowi redundancyjności dla cechy i. Ponieważ jednak dokładne wyliczenie wielkości $\tilde{F}_i$ jest uciążliwe, więc nie będziemy się tu posługiwali tą metodą. Przytoczymy jedynie pewną nierówność prawdziwą dla wspomnianej statystyki testowej $\tilde{F}_i$, a mianowicie

$$\frac{n-J-p+1}{J-1} \frac{U_i}{1+T^2(y_1, ..., y_p) - U_i} \le \tilde{F}_i \le \frac{n-J-p+1}{J-1} U_i . \tag{11.80}$$

Wielkość $\tilde{F}_i$ ma (w przypadku, gdy cecha i jest cechą redundancyjną) rozkład F o stopniach swobody

$$\nu_1 = J - 1, \qquad \nu_2 = n - J - p + 1 . \tag{11.81}$$

Na pytanie, kiedy należy przerwać proces redukcji, a więc kiedy uzyskaliśmy już właściwy kompromis między wielkością miary dyskryminacyjnej, a liczbą cech, można odpowiedzieć uzależniając zakończenie procesu redukcji od tego, jak dalece poszczególne cechy dają istotną zwyżkę miary dyskryminacyjnej.

Jeśli za pomocą redukcji mają być wyeliminowane wszystkie te cechy, których wpływ nie jest statystycznie wystarczająco istotny, to proces redukcji może być na podstawie (11.80) i (11.81) prowadzony aż do momentu, gdy dla wszystkich pozostałych jeszcze cech $y_{i_1}, y_{i_2}, ..., y_{i_{p^*}}$ będzie spełniona nierówność

$$F_{J-1,\, n-J-p^*+1;\, \alpha} < \frac{n-J-p^*+1}{J-1} \frac{U_i}{1+T^2(y_{i_1}, ..., y_{i_p}*) - U_i} \tag{11.82}$$

Istnieje także możliwość prowadzenia procesu redukcji tak długo, aż się przy testowaniu okaże, że zbiór wszystkich wyeliminowanych cech w stosunku do pozostałych cech jest redundancyjny. Wtedy proces redukcji może być przerwany, gdy dla żadnej z pozostałych cech $y_{i_1}, y_{i_2}, \ldots, y_{i_{p^*}}$ nie zachodzi nierówność

$$\frac{n-J-p+1}{(J-1)(p-p^*+1)}\left(T^2(y_1,\ldots,y_p)-T^2(y_{i_1},\ldots,y_{i_p}{}^*)+U_i\right) \leq F_{\nu_1,\nu_2;\alpha} \tag{11.83}$$

Tak jak i we wzorze (11.82) wielkość U_i oznacza tu niezbędność cechy y_i w zbiorze cech $y_{i_1}, y_{i_2}, \ldots, y_{i_{p^*}}$ i ponadto

$$\nu_1 = \begin{cases} \dfrac{(J-1)(p-p^*+1)(n-J-p)}{n-(J-1)(p-p^*+1)-p^*-1}, & \text{gdy } n-(J-1)(p-p^*+1)-p^*-1>0, \\ \infty, & \text{gdy } n-(J-1)(p-p^*+1)-p^*-1\leq 0 \end{cases} \tag{11.84}$$

oraz

$$\nu_2 = n-J-p+1\ . \tag{11.85}$$

W zastosowaniach tych testów istotności przerwanie redukcji będzie w znacznej mierze zależało od wielkości próby n. Im większe n, tym mniej cech zostanie wyeliminowanych.

Przykład 11.

Kończymy już przykład dotyczący schorzenia tarczycowego. Uwzględniając wszystkie 10 cech otrzymamy następujące niezbędności

$$\{U_i\} = (0{,}38 \quad 1{,}31 \quad 0{,}98 \quad 0{,}31 \quad 0{,}03 \quad 1{,}68 \quad 0{,}23 \quad 0{,}12 \quad 1{,}19 \quad 3{,}33)^T$$

Porównanie tych wartości z wartościami miary dyskryminacyjnej poszczególnych cech (wyliczonymi w poprzednich przykładach) pokazuje, że uporządkowanie cech co do rangi według jednowymiarowej zdolności dyskryminacyjnej i uporządkowanie według niezbędności nie są zgodne. Cecha y_9, na przykład, ma wprawdzie bardzo małą jednowymiarową zdolność dyskryminacyjną, ale w kombinacji z innymi cechami, zwłaszcza z y_6 i y_7, okazuje się bardzo pożyteczna. Uwarunkowane jest to między innymi dużymi korelacjami między y_9 i innymi cechami.

Wszystkie przytoczone niezbędności są na podstawie kryterium (11.82) dla $\alpha = 0{,}05$ nieistotnie różne od zera. Konieczna do istotności minimalna wartość niezbędności U_i wynosi 4,47. Brak istotności tkwi w tym, że każda z tych 10 cech jest w dużym stopniu zastępowalna przez pozostałe 9 cech tak, że w konsekwencji specyficzny wpływ jakiejś określonej cechy staje się wątpliwy. Jeśli natomiast przez redukcję wyeliminowane zostaną

cechy zbędne, to zwiększają się perspektywy na to, że niezbędności pozostawionych cech będą istotnie odchylały się od zera.

Posługując się tablicą 11.1 możemy prześledzić cały proces redukcji cech. Proces ten zaczyna się od zbioru 10 cech, a kończy wtedy, gdy pozostaje już tylko jedna cecha (kolumna 2 tab. 11.1). W kolumnie 3 podane są te cechy, które każdorazowo eliminowane są w kolejnym kroku. Kolumna 4 podaje miarę dyskryminacyjną T^2 aktualnie istniejącego zbioru p^* cech; kolumny 5, 6 i 7 zawierają wartość weryfikacyjną $\tilde{F}$ testu wielowymiarowego, odpowiednie stopnie swobody ν_1 i ν_2 oraz krytyczne prawdopodobieństwo błędu α, od którego począwszy przy teście wielowymiarowym zachodzi już istotność. W kolumnie 8 podana jest niezbędność U_i cechy, która ma być wyeliminowana, w kolumnie 9 podana jest wartość weryfikacyjna ze wzoru (11.82)

Tabela 11.1

Pełny proces redukcji cech w schorzeniu tarczycowym

Krok	p^*	cecha reduk.	Miara T^2 pozost. cech	$\tilde{F}$ pozost. cech	Stopnie swobody ν_1 ν_2	α ($*10^{-6}$)	Niezbędność U_i cechy redukow.	Wartość weryf. redukow. cechy	Stopnie swobody ν_1 ν_2	Wartość kryt $F_{\nu_1,\nu_2;0,05}$
1	10	y_5	9,64	5,30	200 11	1930	0,0253	0,0131	2 11	3,98
2	9	y_8	9,62	6,41	66 12	569	0,124	0,0711	2 12	3,89
3	8	y_7	9,49	7,71	38,4 13	159	0,268	0,170	2 13	3,81
4	7	y_4	9,22	9,22	26 14	43,4	0,364	0,258	2 14	3,74
5	6	y_1	8,86	11,1	18,7 15	11,4	0,449	0,358	2 15	3,68
6	5	y_3	8,41	13,5	13,6 16	2,9	1,40	1,40	2 16	3,63
7	4	y_2	7,01	14,9	9,6 17	1,7	1,61	2,13	2 17	3,59
8	3	y_9	5,40	16,2	6,8 18	1,6	1,82	3,59	2 18	3,55
9	2	y_6	3,58	17,0	4,2 19	3,9	1,17	3,26	2 19	3,52
10	1	y_{10}	2,41	24,1	2 20	4,8	2,41	24,1	2 20	3,49

$$\frac{n-J-p^*+1}{J-1}\,\frac{U_i}{1+T^2-u_i},$$

a w kolumnach 10 i 11 pokazane są odpowiednie stopnie swobody, oraz wartość krytyczna F dla $\alpha = 0{,}05$.

Pozostałą przy końcu tego procesu cechą jest cecha y_{10}, którą już poprzednio uznaliśmy jako najlepszą cechę dyskryminacyjną. Uzyskana w procesie redukcji para cech y_6, y_{10} także już poprzednio została uznana za najkorzystniejszą kombinację.

Jeśli opisywana redukcja zostanie doprowadzona tak daleko, aż wyeliminowane zostaną wszystkie te cechy, które według (11.82) nie wnoszą nic istotnego do rozróżnienia populacji, to pozostanie w końcu już tylko trójka cech y_6, y_9, y_{10}. Wpływ każdej z tych trzech cech na diagnozę jest statystycznie pewny, podczas gdy zgodnie z kolumną 9 tablicy 11.1 dla wszystkich obszerniejszych zbiorów cech, które występują w trakcie redukcji, tego rodzaju zapewnienia nie udaje się uzyskać (dla $\alpha = 0{,}05$).

11.3 Wielowymiarowa wieloczynnikowa analiza wariancyjna

Rozpatrywaliśmy dotychczas klasyfikację pojedynczą, gdzie wszystkie rozpatrywane różnice między wartościami średnimi powodowane były przez jeden jedyny czynnik. W licznych jednak przypadkach do wytworzenia się owych różnic między rozpatrywanymi parametrami statystycznymi przyczynia się wiele czynników. W podrozdziale tym omówimy jedynie kompletne i zróżnicowane układy eksperymentalne z klasyfikacją wielokrotną, tzn. takie, w których liczba obserwowanych w każdej komórce wartości jest jednakowa. Z uwagi na brak miejsca ograniczymy się do omówienia jedynie klasyfikacji podwójnej (dwuczynnikowej). W przypadku układów doświadczeń z klasyfikacją potrójną lub więcej krotną i o jednakowym obsadzeniu komórek oraz w przypadku hipotez, które dotyczą całych grup efektów głównych i efektów interakcyjnych, można stosunkowo łatwo otrzymać odpowiednie statystyki testowe według wzoru jaki daje nam klasyfikacja podwójna, o ile tylko będziemy pamiętali, by zawsze wychodzić z odpowiedniego przypadku jednowymiarowego. Wyprowadzenie takie jest zawsze potrzebne — bo chociaż w przypadku wielowymiarowej analizy wariancji zawsze korzystamy z komputera i odpowiedniego oprogramowania — to jednak do odpowiedniej interpretacji wyników musimy wiedzieć na ile wyprowadzone przez nas wzory różnią się od tych podanych w dokumentacji.

11.3.1 Klasyfikacja podwójna, jeden wektor obserwacji na komórkę

Rozpoczniemy od krótkiego przypomnienia przypadku jednowymiarowego, tzn. takiego w którym

$$\dim(y_{jk}) = p = 1 \ .$$

Zakładamy, że czynnik A działa na J poziomach, natomiast czynnik B działa na K poziomach.

Schemat układu eksperymentalnego można przedstawić jak poniżej:

		czynnik B				
		1	2	3	...	K
A	1	y_{11}	y_{12}	y_{13}	...	y_{1K}
	2	y_{21}	y_{22}	y_{23}	...	y_{2K}
	:	.....	.....	.....	.....	.....
	J	y_{J1}	y_{J2}	y_{J3}	...	y_{JK}

(11.86)

Rozpatrujemy oczywiście model liniowy, a zatem przy klasyfikacji podwójnej możemy zapisać następujące założenie:

$$y_{jk} = \mu + \alpha_j + \beta_k + \varepsilon_{jk} \qquad (j = 1, ..., J,\ k = 1, ..., K) \tag{11.87}$$

gdzie zmienne y_{jk} podlegają rozkładowi

$$y_{jk} \approx N(\mu + \alpha_j + \beta_k, \sigma^2)$$

dla wszystkich wartości j oraz k.
Interesują nas hipotezy zerowe postaci:

$$H_{0A}: \alpha_1 = \ldots = \alpha_J,$$
$$H_{0B}: \beta_1 = \ldots = \beta_K. \tag{11.87}$$

Utworzona zgodnie z przyjętymi założeniami tablica analizy wariancji ma postać:

sumy kwadratów	st. swob.	statystyka F
$SS_A = K \sum_j (y_{j.} - y_{..})^2$	$\nu_A = J - 1$	$\frac{\nu_R}{\nu_A} \frac{SS_A}{SS_R} \approx F_{\nu_A, \nu_R}$
$SS_B = J \sum_k (y_{.k} - y_{..})^2$	$\nu_B = K - 1$	$\frac{\nu_R}{\nu_B} \frac{SS_B}{SS_R} \approx F_{\nu_B, \nu_R}$
$SS_R = \sum_j \sum_k (y_{jk} - y_{j.} - y_{.k} + y_{..})^2$	$\nu_R = (J - 1)(K - 1)$	

Wielkość SS_R / ν_R jest nieobciążoną oceną wariancji σ^2 błędu doświadczenia. Ponadto zachodzą zależności:

$$y_{j.} = \frac{1}{K}\sum_{k} y_{jk}\,, \qquad y_{.k} = \frac{1}{J}\sum_{j} y_{jk}\,, \qquad y_{..} = \sum_{j}\sum_{k} y_{jk}\,, \tag{11.88}$$

Rozpatrzmy teraz przypadek wielowymiarowy, tzn. taki dla którego zachodzi

$$\dim(y_{jk}) = p > 1\ .$$

Wielkości obserwacyjne są realizacjami p-wymiarowych wektorów losowych o rozkładach normalnych. Schemat klasyfikacji podwójnej odpowiada przypadkowi jednowymiarowemu z tym, że obecnie zamiast komórki (j, k) jest wektor obserwacji y_{jk}, który obejmuje p mierzonych cech badanego obiektu. Równaniem modelu jest

$$y_{jk} = \mu + \alpha_j + \beta_k + \varepsilon_{jk} \qquad (j = 1, ..., J,\ k = 1, ..., K)\ , \tag{11.89}$$

gdzie tym razem μ, α_j, β_k są odpowiednimi p-wymiarowymi wektorami a również p-wymiarowy wektor ε_{jk} podlega rozkładowi:

$$\varepsilon_{jk} \approx N(0, \Sigma) \text{ dla wszystkich } j \text{ oraz } k\ . \tag{11.90}$$

Chcemy zweryfikować następujące hipotezy zerowe:

$$H_{0A} : \alpha_1 = ... = \alpha_J\ , \tag{11.91}$$

$$H_{0B} : \beta_1 = ... = \beta_K\ . \tag{11.92}$$

Tak jak to już opisywaliśmy poprzednio, w celu uzyskania statystyki testowej musimy znaleźć macierze H_A, H_B i G, które wynikają bezpośrednio z jednowymiarowej tablicy analizy wariancji jako uogólnienie wyrażeń SS_A, SS_B i SS_R:

$$H_A = K\sum_{j}(y_{j.} - y_{..})(y_{j.} - y_{..})^T\ , \tag{11.93}$$

$$H_B = J\sum_{k}(y_{.k} - y_{..})(y_{.k} - y_{..})^T\ , \tag{11.94}$$

$$G = \sum_{j}\sum_{k}(y_{jk} - y_{j.} - y_{.k} + y_{..})(y_{jk} - y_{j.} - y_{.k} + y_{..})^T \tag{11.95}$$

przy czym wielkości $y_{j.}$, $y_{.k}$ oraz $y_{..}$ wynikają z (11.88), gdy zamiast y_{jk} podstawi się wektory y_{jk}.

Jako statystyki testowe przyjmiemy

$$\frac{\nu_R - p + 1}{\nu_A p} \; tr\,(H_A G^{-1}) \quad \text{dla hipotezy } H_{0A} \tag{11.96}$$

$$\frac{\nu_R - p + 1}{\nu_B p} \; tr\,(H_B G^{-1}) \quad \text{dla hipotezy } H_{0B} \tag{11.97}$$

Macierz

$$S = \frac{1}{\nu_R} G \tag{11.98}$$

jest nieobciążoną oceną macierzy kowariancyjnej Σ naszego modelu liniowego. Przekształcając wzory (11.96) i (11.97) otrzymujemy

$$tr\,(H_A G^{-1}) = \frac{K}{\nu_R} \sum_j (y_{j.} - y_{..})^T S^{-1} (y_{j.} - y_{..}) \tag{11.99}$$

oraz

$$tr\,(H_B G^{-1}) = \frac{J}{\nu_R} \sum_k (y_{.k} - y_{..})^T S^{-1} (y_{.k} - y_{..}) \tag{11.100}$$

Zatem do weryfikacji naszych hipotez otrzymujemy następujące statystyki testowe:

Hipoteza H_{0A} : $\alpha_1 = \ldots = \alpha_J$.

Wielkość

$$\begin{aligned} \tilde{F} &= \frac{(J-1)(K-1) - p + 1}{(J-1)\,p} \; tr\,(H_A G^{-1}) = \\ &= \frac{[(J-1)(K-1) - p + 1]\,K}{(J-1)^2 (K-1)\,p} \sum_{j=1}^{J} (y_{j.} - y_{..})^T S^{-1} (y_{j.} - y_{..}) \end{aligned} \tag{11.101}$$

ma w przybliżeniu rozkład F o ν_1 i ν_2 stopniach swobody, przy czym

$$\nu_1 = \begin{cases} \dfrac{(J-1)\,p\,[(J-1)(K-1) - p]}{(J-1)(K-p) - 1}, & \text{gdy } (J-1)(K-1) - p > 0 \\ \infty & \text{gdy } (J-1)(K-1) - p \geq 0 \end{cases} \tag{11.102}$$

$$\nu_2 = (J-1)(K-1) - p + 1 \tag{11.103}$$

Przy z góry zadanym prawdopodobieństwie α hipotezę H_{0A} odrzucamy, o ile

$$\tilde{F} > F_{\nu_1, \nu_2; \alpha} \tag{11.104}$$

Drugą z naszych hipotez jest zgodnie z (11.92) hipoteza:

Hipoteza H_{0B} : $\beta_1 = \ldots = \beta_K$.

Wielkość

$$\begin{aligned} \tilde{F} &= \frac{(J-1)(K-1) - p + 1}{(J-1)\,p}\, tr\,(H_B\, G^{-1}) = \\ &= \frac{[(J-1)(K-1) - p + 1]\,J}{(K-1)^2\,(J-1)\,p} \sum_{k=1}^{K} (y_{.k} - y_{..})^T S^{-1} (y_{.k} - y_{..}) \end{aligned} \tag{11.105}$$

ma w przybliżeniu rozkład F o ν_1 i ν_2 stopniach swobody, przy czym

$$\nu_1 = \begin{cases} \dfrac{(K-1)\,p\,[(J-1)(K-1) - p]}{(K-1)(J-p) - 1}, & \text{gdy } (K-1)(J-p) - 1 > 0 \\ \infty & \text{gdy } (K-1)(J-1) - 1 \le 0 \end{cases} \tag{11.106}$$

$$\nu_2 = (J-1)(K-1) - p + 1 \tag{11.107}$$

Przy z góry zadanym prawdopodobieństwie α hipotezę H_{0B} odrzucamy, o ile

$$\tilde{F} > F_{\nu_1, \nu_2; \alpha} \tag{11.108}$$

11.3.2 Klasyfikacja podwójna, *m* wektorów obserwacji na komórkę

Różnica w stosunku do klasyfikacji podwójnej rozważanej w poprzednim punkcie tkwi jedynie w tym, że obecnie w każdej komórce (oznaczonej (j, k)) znajduje się jednakowa liczba $m > 1$ wektorów obserwacji

$$y_{jk1}, y_{jk2}, \ldots, y_{jkn} ,$$

które zawierają po p cech. W związku z tym w równaniu modelu powinien zostać uwzględniony efekt interakcji j-tego poziomu czynnika A z k-tym poziomem czynnika B. Interakcja możliwa jest naturalnie i dla $m = 1$, ale dopiero dla $m > 1$ daje się analizować. Tak jak i poprzednio rozpoczniemy od przypomnienia przypadku jednowymiarowego w którym

$$\dim (y_{jkl}) = p = 1 \ .$$

Równanie modelu ma postać:

$$y_{jkl} = \mu + \alpha_j + \beta_k + (\alpha\,\beta)_{jk} + \varepsilon_{jkl} \qquad j = 1, ..., J,\ k = 1, ..., K,\ l = 1, ..., m \tag{11.109}$$

przy czym

$$y_{jkl} = N\,(\mu + \alpha_j + \beta_k + (\alpha\,\beta)_{jk}\,,\ \sigma^2) \tag{11.110}$$

oraz

$$\sum_{j=1}^{J} \alpha_j = 0; \qquad \sum_{k=1}^{K} \beta_k = 0; \qquad \sum_{j=1}^{J} (\alpha\,\beta)_{jk} = 0 \quad \text{dla wszystkich } k$$

oraz

$$\sum_{k=1}^{K} (\alpha\,\beta)_{jk} = 0 \quad \text{dla wszystkich } j.$$

Stosowna tablica analizy wariancji zamieszczona została w tabeli 11.2.
Użyte w niej oznaczenia są następujące:

$$y_{jk.} = \frac{1}{m} \sum_{l} y_{jkl}\,, \qquad y_{j..} = \frac{1}{m\,K} \sum_{k} \sum_{l} y_{jkl}\,,$$

$$y_{.k.} = \frac{1}{m\,J} \sum_{j} \sum_{l} y_{jkl}\,, \qquad y_{...} = \frac{1}{m\,J\,K} \sum_{j} \sum_{k} \sum_{l} y_{jkl}\,.$$

Wartość SS_R / ν_R jest nieobciążoną oceną wariancji błędu eksperymentalnego σ^2.

Rozpatrzmy teraz przypadek wielowymiarowy, dla którego zachodzi

$$\dim (y_{jkl}) = p > 1 \ .$$

Równanie modelu liniowego dla l-tego wektora wartości obserwacyjnych w komórce (j, k) ma postać jak poprzednio:

Tabela 11.2

Tablica analizy wariancji w przypadku dwuczynnikowym jednowymiarowym

Sumy kwadratów	Stopnie swobody	Statystyka testowa
$SS_A = mK \sum_j (y_{j..} - y_{...})^2$	$\nu_A = J - 1$	dla H_{0A}: $\frac{\nu_R}{\nu_A} \frac{SS_A}{SS_R} \approx F_{\nu_A, \nu_R}$
$SS_B = mJ \sum_k (y_{.k.} - y_{...})^2$	$\nu_B = K - 1$	dla H_{0B}; $\frac{\nu_R}{\nu_B} \frac{SS_B}{SS_R} \approx F_{\nu_B, \nu_R}$
$SS_{A\times B} = m \sum_j \sum_k (y_{jk.} - y_{j..} - y_{.k.} + y_{...})^2$	$\nu_{A\times B} = (J-1)(K-1)$	dla $H_{0A\times B}$; $\frac{\nu_R}{\nu_{A\times B}} \frac{SS_{A\times B}}{SS_R} \approx F_{\nu_{A\times B}, \nu_R}$
$SS_R = \sum_j \sum_k \sum_l (y_{jkl} - y_{jk.})^2$	$\nu_R = JK(m-1)$	

$$y_{jkl} = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk} + \varepsilon_{jkl} \qquad j = 1, ..., J,\ k = 1, ..., K,\ l = 1, ..., m \tag{11.111}$$

z tym, że teraz wielkości α_j, β_k, $(\alpha\beta)_{jk}$ są p-wymiarowymi wektorami spełniającymi warunki:

$$\sum_{j=1}^{J} \alpha_j = 0; \qquad \sum_{k=1}^{K} \beta_k = 0; \qquad \sum_{j=1}^{J} (\alpha\beta)_{jk} = 0$$

dla wszystkich k oraz

$$\sum_{k=1}^{K} (\alpha\beta)_{jk} = 0$$

dla wszystkich j. Dla wektorów błędów doświadczenia zakłada się, że

$$\varepsilon_{jkl} \approx N(0, \Sigma)$$

dla wszystkich j, k oraz l. Weryfikować będziemy następujące hipotezy zerowe:

$$H_{0A\times B}: (\alpha\beta)_{11} = \ldots = (\alpha\beta)_{JK} = 0,$$
$$H_{0A}: \alpha_1 = \ldots = \alpha_j = 0 \qquad (11.112)$$
$$H_{0B}: \beta_1 = \ldots = \beta_k = 0$$

Do tego potrzebne jest obliczenie macierzy H_A, H_B $H_{A\times B}$ i G, które mają następujące postacie:

$$H_A = mK \sum_j (y_{j..} - y_{...})(y_{j..} - y_{...})^T, \qquad (11.113)$$

$$H_B = mJ \sum_k (y_{.k.} - y_{...})(y_{.k.} - y_{...})^T, \qquad (11.114)$$

$$H_{A\times B} = m \sum_j \sum_k (y_{jk.} - y_{j..} - y_{.k.} + y_{...})(y_{jk.} - y_{j..} - y_{.k.} + y_{...})^T. \qquad (11.115)$$

$$G = \sum_j \sum_k \sum_l (y_{jkl} - y_{jk.})(y_{jkl} - y_{jk.})^T \qquad (11.116)$$

Wielkości $y_{jk.}$, $y_{j..}$, $y_{.k.}$ i $y_{...}$ wylicza się tak jak wielkości w tab. 11.2 z tą różnicą, że w miejsce y_{jkl} podstawia się p-wymiarowe wektory y_{jkl}.
Jako statystyki testowe stosujemy

$$\frac{\nu_R - p + 1}{\nu_{A\times B}\, p}\, tr(H_{A\times B} G^{-1}) \quad \text{dla hipotezy } H_{0A\times B} \qquad (11.117)$$

$$\frac{\nu_R - p + 1}{\nu_A\, p}\, tr(H_A G^{-1}) \quad \text{dla hipotezy } H_{0A} \qquad (11.118)$$

$$\frac{\nu_R - p + 1}{\nu_B\, p}\, tr(H_B G^{-1}) \quad \text{dla hipotezy } H_{0B} \qquad (11.119)$$

Tak jak i poprzednio macierz

$$S = \frac{1}{\nu_R} G$$

jest nieobciążoną oceną macierzy kowariancyjnej Σ.
Zachodzi również:

$$tr\,(H_{A\times B}\,G^{-1}) = \\ = \frac{m}{JK\,(m-1)} \sum_j \sum_k (y_{jk\cdot} - y_{j\cdot\cdot} - y_{\cdot k\cdot} + y_{\cdot\cdot\cdot})^T\, S^{-1}\, (y_{jk\cdot} - y_{j\cdot\cdot} - y_{\cdot k\cdot} + y_{\cdot\cdot\cdot}) \tag{11.120}$$

$$tr\,(H_A\,G^{-1}) = \frac{m}{J\,(m-1)} \sum_j (y_{j\cdot\cdot} - y_{\cdot\cdot\cdot})^T\, S^{-1}\, (y_{j\cdot\cdot} - y_{\cdot\cdot\cdot}) \tag{11.121}$$

$$tr\,(H_B\,G^{-1}) = \frac{m}{K\,(m-1)} \sum_k (y_{\cdot k\cdot} - y_{\cdot\cdot\cdot})^T\, S^{-1}\, (y_{\cdot k\cdot} - y_{\cdot\cdot\cdot}) \tag{11.122}$$

Zatem do weryfikacji hipotez zerowych postaci (11.112) otrzymujemy następujące statystyki testowe:

Hipoteza $H_{0A\times B} : (\alpha\beta)_{11} = \ldots = (\alpha\beta)_{JK} = 0$.

Statystyka testowa

$$\tilde{F} = \frac{[JK\,(m-1) - p + 1]\,m}{(J-1)\,(K-1)\,p\,J\,K\,(m-1)} \cdot \\ \cdot \sum_j \sum_k (y_{jk\cdot} - y_{j\cdot\cdot} - y_{\cdot k\cdot} + y_{\cdot\cdot\cdot})^T\, S^{-1}\, (y_{jk\cdot} - y_{j\cdot\cdot} - y_{\cdot k\cdot} + y_{\cdot\cdot\cdot}) \tag{11.123}$$

ma w przybliżeniu rozkład F o ν_1 oraz ν_2 stopniach swobody, gdzie

$$\nu_1 = \begin{cases} \dfrac{(J-1)\,(K-1)\;p\,[JK\,(m-1)-p]}{JK\,(m-1) - (p-1)\,(J-1)\,(K-1) - 1}, & \text{gdy mianownik} > 0 \\ \infty & \text{gdy mianownik} \le 0 \end{cases} \tag{11.124}$$

$$\nu_2 = JK(m-1) - p + 1 \tag{11.125}$$

Przy danym prawdopodobieństwie błędu α hipotezę $H_{0A\times B}$ odrzuca się, o ile

$$\tilde{F} > F_{\nu_1, \nu_2; \alpha} \tag{11.126}$$

Następną hipotezą jest:

Hipoteza $H_{0A} : \alpha_1 = \ldots = \alpha_J = 0$.

Statystyka testowa

$$\tilde{F} = \frac{(JK(m-1)-p+1)\,m}{(J-1)\,p\,J(m-1)} \cdot \sum_{j} (y_{j..} - y_{...})^T S^{-1} (y_{j..} - y_{...}) \tag{11.127}$$

ma w przybliżeniu rozkład F o ν_1 oraz ν_2 stopniach swobody, gdzie

$$\nu_1 = \begin{cases} \dfrac{JK(m-1)-p)\,(K-1)\,p}{JK(m-1)-(J-1)\,(p-1)-1}, & \text{gdy mianownik} > 0 \\ \infty & \text{gdy mianownik} \le 0 \end{cases} \tag{11.128}$$

$$\nu_2 = JK(m-1) - p + 1 \tag{11.129}$$

Przy danym prawdopodobieństwie błędu α hipotezę H_{0A} odrzuca się, o ile

$$\tilde{F} > F_{\nu_1, \nu_2; \alpha} \tag{11.130}$$

Ostatnia już testowana hipoteza to:

Hipoteza H_{0B} : $\beta_1 = \ldots = \beta_K = 0$.

Statystyka testowa

$$\tilde{F} = \frac{(JK(m-1)-p+1)\,m}{(K-1)\,p\,K(m-1)} \cdot \sum_{k} (y_{.k.} - y_{...})^T S^{-1} (y_{.k.} - y_{...}) \tag{11.131}$$

ma w przybliżeniu rozkład F o ν_1 oraz ν_2 stopniach swobody, gdzie

$$\nu_1 = \begin{cases} \dfrac{(JK(m-1)-p)\,(K-1)\,p}{JK(m-1)-(K-1)\,(p-1)-1}, & \text{gdy mianownik} > 0 \\ \infty & \text{gdy mianownik} \le 0 \end{cases} \tag{11.132}$$

$$\nu_2 = JK(m-1) - p + 1 \tag{11.133}$$

Przy danym prawdopodobieństwie błędu α hipotezę H_{0B} odrzuca się, o ile

$$\tilde{F} > F_{\nu_1, \nu_2; \alpha} \tag{11.134}$$

Zgodnie z poczynioną na początku rozdziału uwagą omówimy jeszcze pokrótce założenia wielowymiarowej analizy wariancji. Są one następujące:

— wielowymiarowy rozkład normalny wektorów wartości obserwacyjnych,

— jednakowość macierzy kowariancyjnych,
— stochastyczna niezależność wektorów wartości obserwacyjnych,
— istnienie wartości pomiarowych wszystkich p cech w każdym badanym obiekcie.

Założenia te rzadko kiedy są ściśle spełnione. Nie oznacza to jednak, że nie możemy stosować metod analizy wariancyjnej i dyskryminacyjnej. Nie należy sądzić, że w wielowymiarowej analizie wariancji każde odstępstwo od założeń musi prowadzić do dużych zniekształceń w praktycznym wnioskowaniu. Przeciwnie, główne wyniki analizy powinny jednak pozostać prawdziwe.

Niespełnienie wspomnianych założeń może powodować to, że ten czy inny z opisywanych testów istotności przeprowadzany na poziomie istotności np. $\alpha = 0{,}05$, ma przy uwzględnieniu rzeczywiście prawdziwego, ale nieznanego, rozkładu statystyki testowej nieco inny poziom istotności, np. $\alpha = 0{,}02$. Ponieważ z reguły przyjęcie określonego poziomu istotności jest arbitralne, zatem takie jego „przesunięcie" nie zmieni otrzymanych wyników.

W większości zastosowań możemy również przeprowadzić wielowymiarową analizę, o ile mamy do czynienia z cechami ciągłymi, nie analizując bliżej wielowymiarowego rozkładu wektorów cech. Niejednokrotnie dobre rezultaty osiąga się poddając określone cechy wstępnej transformacji, np. logarytmicznej.

Również niespełnienie warunku o równości macierzy kowariancji nie prowadzi na ogół do żadnych poważniejszych zakłóceń w wynikach. Dodatkowo w przypadku wielowymiarowej analizy wariancji przy niejednakowych kowariancjach istnieją pewne założenia przybliżone.

12. REGRESJA WIELOKROTNA

W licznych badaniach biometrycznych obserwuje się równocześnie wiele cech charakteryzujących populację biologiczną. Cechy te mogą być (i na ogół są) wzajemnie zależne, w związku z tym istotne jest określenie siły tej zależności oraz jej wykorzystanie do przewidywania jednej cechy na podstawie innej. W pierwszej części skryptu rozpatrywaliśmy zależność między dwiema cechami (zależną i niezależną), opisując ten związek funkcją liniową — prostą regresji. Gdybyśmy mieli zespół trzech zmiennych Y, X_1, X_2 i chcielibyśmy określać wartości cechy Y na podstawie X_1 lub X_2, to można to zrobić konstruując dwie funkcje regresji

$$Y = f_1(X_1) \qquad \text{oraz} \qquad Y = f_2(X_2)$$

i wybierając tę z nich, która jest dokładniejsza. Wydaje się jednak intuicyjnie jasne, że łączne wnioskowanie o cesze Y na podstawie wartości obu zmiennych X_1 i X_2 powinno dać wyniki lepsze (a już na pewno nie gorsze) niż rozpatrywanie pojedynczych związków.

12.1 Równanie regresji wielokrotnej

Rozpatrujemy zespół cech Y, X_1, X_2, ..., X_p o których zakładamy, że są określone na wszystkich elementach populacji. Liniowy związek między wartością cechy i-tego elementu populacji, tzn. y_i, a wartościami pozostałych cech tego elementu, tzn. x_i, definiujemy za pomocą równości:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + e_i \tag{12.1}$$

Analizując powyższą zależność możemy stwierdzić (jest to uproszczenie), że dzieli ona wartość y_i na trzy części:

— część β_0 wspólną dla wszystkich elementów populacji,

— części $\beta_i x_{ij}$ wynikające z „czystego" wpływu danej cechy,

— część e_i wynikającą z wpływu czynników losowych, specyficznych dla y_i, tzn. takich, które nie są związane z pozostałymi cechami.

Identycznie jak w przypadku dwóch cech, chcemy tak dobrać współczynniki β_i, aby udział czynnika losowego e_i był jak najmniejszy. Prowadzi to do konstrukcji funkcji regresji $f(x_1, x_2, ..., x_p)$, dla której wartość oczekiwana kwadratu odchylenia Y od f osiąga minimum, tzn.:

$$E\,[Y - f(x_1, x_2, ..., x_p)]^2 = \min\ .$$

Funkcji regresji można użyć do przewidywania (prognozy) wartości zmiennej losowej Y gdy znane są wartości zmiennych losowych $X_1, X_2, ..., X_p$. Równanie służące do takiego przewidywania nazywa się równaniem regresji wielokrotnej i ma ono postać:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p \tag{12.2}$$

Współczynniki β_i noszą nazwę cząstkowych współczynników regresji.

Wartości współczynników β_i odnoszą się do populacji i na ogół nie są znane, lecz podlegają oszacowaniu na podstawie próby. Oznaczmy zatem przez n liczebność próby, natomiast przez b_0, b_1, b_2, ... b_p estymatory parametrów β_0, β_1, β_2, ... β_p. Równanie regresji wielokrotnej przyjmuje wtedy postać:

$$y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p$$

Jeżeli przez φ oznaczymy sumę kwadratów odchyleń funkcji regresji od Y:

$$\varphi = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = (y_i - \beta_0 - \beta_1 x_1 - \ldots - \beta_p x_p)^2$$

to współczynniki b_i wyznaczymy rozwiązując układ równań:

$$\frac{\delta \varphi}{\delta \beta_i} = 0, \qquad i = 1, 2, ..., p$$

Pochodne cząstkowe mają postać:

$$\frac{\delta \varphi}{\delta \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} - \ldots - \beta_p x_{pi})$$

$$\frac{\delta \varphi}{\delta \beta_j} = -2 \sum_{i=1}^{n} x_{ji} (y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} - \ldots - \beta_p x_{pi}), \qquad j = 1, 2, ..., p$$

Po przyrównaniu pochodnych do zera, zastąpieniu β_j ich estymatorami b_j, podzieleniu obu stron równań przez 2 i przeniesieniu wyrazów nie zawierających niewiadomych b_j

na prawą stronę, otrzymamy układ równań, który w postaci macierzowej można zapisać jak poniżej:

$$\boldsymbol{A}\ \boldsymbol{b} = \boldsymbol{w} \tag{12.3}$$

gdzie

$\boldsymbol{A} = [a_{jk}]$, $j = 0, \ldots, p$, $k = 0, \ldots, p$ — jest macierzą współczynników przy niewiadomych $b_0, b_1, b_2, \ldots b_p$,

$$a_{jk} = \sum_{i=1}^{n} x_{ji}\, x_{ki} \quad \text{przy czym } x_{0i} = 1 \quad \text{dla} \quad i = 1, 2, \ldots, n$$

$\boldsymbol{b} = [b_0, b_1, \ldots, b_p]^T$ — jest wektorem niewiadomych,

$\boldsymbol{w} = [w_j]^T$ — jest wektorem wyrazów wolnych,

$$w_j = \sum_{i=1}^{n} x_{ji}\, y_i, \quad \text{przy czym } x_{0i} = 1 \quad \text{dla} \quad i = 1, 2, \ldots, n$$

Rozwiązaniem równania (12.3) jest wektor

$$\boldsymbol{b} = \boldsymbol{A}^{-1}\ \boldsymbol{w}$$

Układ równań określony wzorem (12.3) nazywa się układem równań normalnych, rozwiązaniem jest wektor wartości współczynników b_i. Znając te współczynniki można z kolei przewidywać wartość zmiennej zależnej y na podstawie zaobserwowanych wartości zmiennych $x_1, x_2, \ldots, x_p$:

$$y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p \tag{12.4}$$

Układ równań normalnych można też przedstawić w nieco innej postaci. Jeżeli pierwsze równanie układu podzielimy obustronnie przez n, to otrzymamy

$$b_0 + b_1 \overline{x}_1 + b_2 \overline{x}_2 + \ldots + b_p \overline{x}_p = \overline{y}$$

gdzie

$\overline{x}_j = \frac{1}{n} \sum_i x_{ji}$ — oznacza średnią arytmetyczną cechy X_j,

$\overline{y} = \frac{1}{n} \sum_i y_i$ — oznacza średnią arytmetyczną cechy Y.

Z równania tego wyliczamy

$$b_0 = \bar{y} - b_1 \bar{x}_1 - b_2 \bar{x}_2 - \ldots - b_p \bar{x}_p$$

i po podstawieniu do (12.4) otrzymujemy:

$$y - \bar{y} = b_1 (x_1 - \bar{x}_1) + b_2 (x_2 - \bar{x}_2) + \ldots + b_p (x_p - \bar{x}_p) \tag{12.5}$$

Odpowiadający temu równaniu regresji układ równań normalnych jest bardzo zbliżony do układu (12.3), a różnice polegają na tym, że nie występuje w nim pierwszy wiersz i pierwsza kolumna, a w pozostałych wierszach i kolumnach symbole x_{ji} oraz y_i są zastąpione odchyleniami od średnich. Można go zatem zapisać jako:

$$\boldsymbol{S}\ \boldsymbol{b} = \boldsymbol{s}_y \tag{12.6}$$

gdzie

$$\boldsymbol{S} = \begin{bmatrix} S_{11} & S_{12} & \ldots & S_{1p} \\ S_{12} & S_{22} & \ldots & S_{2p} \\ \ldots & \ldots & \ldots & \ldots \\ S_{1p} & S_{2p} & \ldots & S_{pp} \end{bmatrix}, \qquad \boldsymbol{b} = \begin{bmatrix} b_1 \\ b_2 \\ . \\ . \\ b_p \end{bmatrix}, \qquad \boldsymbol{s}_y = \begin{bmatrix} s_{1y} \\ s_{2y} \\ . \\ . \\ s_{py} \end{bmatrix},$$

oraz

$$S_{jj} = S_{x_j} S_{x_j} = \sum_{i=1}^{n} (x_{ji} - \bar{x}_j)^2$$

$$S_{jk} = S_{x_j x_k} = \sum_{i=1}^{n} (x_{ji} - \bar{x}_j)(x_{ki} - \bar{x}_k)$$

$$S_{jy} = S_{x_j y} = \sum_{i=1}^{n} (x_{ji} - \bar{x}_j)(y_i - \bar{y})$$

Każde z równań układu (12.6) możemy z kolei podzielić przez $(n - 1)$ i wtedy zamiast sum kwadratów odchyleń wystąpią wariancje, zamiast zaś sum iloczynów odchyleń — kowariancje. W ten sposób otrzymamy ostateczną wersję układu równań normalnych:

$$\boldsymbol{C}\ \boldsymbol{b} = \boldsymbol{c}_y \tag{12.7}$$

gdzie

$$\boldsymbol{C}=\begin{bmatrix} s_1^2 & s_{12} & \cdots & s_{1p} \\ s_{12} & s_2^2 & \cdots & s_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ s_{1p} & s_{2p} & \cdots & s_p^2 \end{bmatrix}, \qquad \boldsymbol{b}=\begin{bmatrix} b_1 \\ b_2 \\ . \\ . \\ b_p \end{bmatrix}, \qquad \boldsymbol{c}_y=\begin{bmatrix} s_{1y} \\ s_{2y} \\ . \\ . \\ s_{py} \end{bmatrix},$$

oraz

$$s_j^2=\frac{S_{yy}}{n-1}, \qquad s_{jk}=\frac{S_{jk}}{n-1}, \qquad s_{jy}=\frac{S_{jy}}{n-1},$$

przy czym

$$\boldsymbol{C}=\frac{1}{n-1}\boldsymbol{S}, \qquad \boldsymbol{c}_y=\frac{1}{n-1}\boldsymbol{s}_y$$

Macierz $\boldsymbol{C}$ to oczywiście macierz kowariancji — jej elementami diagonalnymi są wariancje poszczególnych zmiennych niezależnych, a elementami pozadiagonalnymi kowariancje tych zmiennych. Macierz kowariancji jest macierzą symetryczną.

Rozważając regresję prostoliniową między dwiema zmiennymi (tzn. zmienną zależną i zmienną niezależną) stwierdziliśmy, że współczynnik regresji (oznaczany symbolem b) określa o ile — przeciętnie — zmieni się wartość zmiennej zależnej y, gdy wartość zmiennej niezależnej zmieni się o jedną jednostkę.

Interpretacja cząstkowych współczynników regresji jest analogiczna, a więc cząstkowy współczynnik regresji b_j określa o ile przeciętnie zmieni się wartość zmiennej y, gdy wartość zmiennej x_j zmieni się o jedną jednostkę, a wartości pozostałych zmiennych niezależnych pozostaną niezmienione.

UWAGA: Łączna zmiana y wynikająca ze zmiany wszystkich zmiennych niezależnych nie jest zwykłą sumą wynikającą ze zmian poszczególnych x_j traktowanych niezależnie od siebie. Oznacza to, że za b_j nie można przyjmować zwykłych współczynników regresji określonych dla par (y, x_j); współczynniki b_j muszą być określane jednocześnie z uwzględnieniem zależności między wszystkimi zmiennymi.

Przykład 1.

Rozpatrzmy równanie regresji wielokrotnej wiążące ze sobą cechy $x1$, $x2$, $x3$, $x4$ pacjentów należących do 1-szej klasy (por. Dodatek 1). Jako zmienną zależną przyjmiemy stężenie izotopu jodu we krwi po 48 godzinach po podaniu preparatu — czyli cechę $x4$. Szukamy funkcji regresji prostoliniowej $x4 = f(x1, x2, x3)$. Po rozwiązaniu układu równań normalnych otrzymujemy zależność:

$$x4 = -1{,}826 - 0{,}079x1 - 0{,}397x2 + 1{,}380x3$$

Dla porównania, gdybyśmy chcieli przewidywać $x4$ na podstawie tylko jednej cechy, to współczynniki regresji prostej byłyby równe:

$$b_{x4x1} = 0{,}380,\ b_{x4x2} = 0{,}755,\ b_{x4x3} = 1{,}004$$

12.2 Rozwiązywanie układu równań normalnych

Z podanych powyżej wyprowadzeń wynika, że dla znalezienia ocen parametrów β_0, β_1, ... ,β_p należy rozwiązać układ równań normalnych ze względu na $p + 1$ niewiadomych. Rozwiązanie takie jest jednoznaczne, jeżeli wyznacznik macierzy $\boldsymbol{A}$ (lub $\boldsymbol{S}$ albo $\boldsymbol{C}$) jest różny od zera.

Z reguły do rozwiązywania układu równań normalnych stosuje się pewne metody wypracowane na gruncie analizy numerycznej, a pozwalające na uproszczenie obliczeń oraz zapewniające dobrą stabilność numeryczną.

Oceny parametrów β_0, β_1, ... ,β_p można np. wyznaczyć bez odwracania macierzy, a mianowicie stosując tzw. *metodę Doolittle'a*. Idea tej metody polega na takim przekształceniu symetrycznego układu równań liniowych, by uzyskać układ trójkątny, tj. układ, którego wyznacznik główny ma same zera poniżej głównej przekątnej i jedynki na niej. Przekształcanie to polega na mnożeniu równań układu przez pewne liczby oraz dodawaniu i odejmowaniu tych równań. Rozwiązanie układu trójkątnego jest bezpośrednie.

Inną szeroko stosowaną metodą jest *metoda Gaussa-Jordana*, która wykorzystywana jest do rozwiązywania układu równań postaci (12.6), tzn. układu w którym występują sumy kwadratów i iloczynów odchyleń. Dane wyjściowe składają się z elementów macierzy $\boldsymbol{S}$, wektora s_y, oraz sumy kwadratów odchyleń zmiennej y, czyli S_{yy}. Oznaczmy przez $\boldsymbol{U}$ macierz utworzoną z wymienionych elementów według schematu:

$$\boldsymbol{U} = [u_{ij}] = \begin{bmatrix} \boldsymbol{S} & s_y \\ s_y^T & S_{yy} \end{bmatrix} = \left[\begin{array}{ccc:c} S_{11} & \cdots & S_{1p} & S_{1y} \\ \cdots & \cdots & \cdots & \cdots \\ S_{1p} & \cdots & S_{pp} & S_{py} \\ \hdashline S_{1y} & \cdots & S_{py} & S_{yy} \end{array}\right]$$

Jest to macierz symetryczna o $p + 1$ wierszach i kolumnach. Istota omawianej metody polega na wprowadzaniu do macierzy $\boldsymbol{U}$ kolejnych zmiennych niezależnych, przy czym wprowadzanie to będzie się wiązać z wykonaniem odpowiednich przekształceń. Po wprowadzeniu pierwszej zmiennej (np. x_1) i związanych z tym przekształceniach, wprowadzamy

na tej samej zasadzie następne — aż do ostatniej, przy czym kolejność wprowadzania zmiennych jest dowolna.

Niech r będzie numerem wprowadzanej zmiennej ($r = 1, 2, ..., p$), natomiast u'_{ij} będą elementami macierzy uzyskanej po wprowadzeniu nowej zmiennej. Wartości u'_{ij} uzyskuje się z następujących przekształceń wykonywanych w podanej kolejności:

1. $u'_{rr} = -\frac{1}{u_{rr}}$,

2. $u'_{ir} = u'_{ri} = u_{ir} u'_{rr}$ dla $i \neq r$,

3. $u'_{ij} = u'_{ji} = u_{ij} + u_{ir} u'_{rj}$ dla $i \neq r$, $j \neq r$.

Pierwszy ze wzorów przekształca element diagonalny macierzy U odpowiadający wprowadzanej zmiennej, drugi przekształca inne elementy z wiersza (kolumny) odpowiadającego wprowadzanej zmiennej, wreszcie trzeci przekształca pozostałe elementy macierzy.

Wprowadzanie pierwszej zmiennej jest odpowiednikiem rozpatrywania zależności zmiennej y tylko od zmiennej x_1:

$$y - \bar{y} = b_1 (x_1 - \bar{x}_1) \ .$$

Po zakończeniu przekształceń związanych z wprowadzaniem pierwszej zmiennej, w ostatnim (p + 1)-szym wierszu w kolumnie odpowiadającej tej zmiennej, a więc w pierwszej, wystąpi ze znakiem przeciwnym wartość współczynnika regresji b, natomiast w ostatniej (p + 1)-szej kolumnie tego wiersza — resztowa suma kwadratów zmiennej y, tzn. suma kwadratów odchyleń zmiennej y, jaka pozostanie po wyeliminowaniu wpływu zmiennej x_1[1].

Wprowadzanie następnej zmiennej (x_2) oznacza, że rozpatrywana jest zależność y od dwóch zmiennych:

$$y - \bar{y} = b_1 (x_1 - \bar{x}_1) + b_2 (x_2 - \bar{x}_2) \ .$$

Po zakończeniu zdefiniowanych powyżej przekształceń, w ostatnim wierszu, w kolumnach odpowiadających wprowadzonym dotychczas zmiennym występują (ze znakiem przeciwnym) współczynniki regresji b_1 i b_2, natomiast w kolumnie ostatniej — resztowa suma kwadratów pozostała po wyeliminowaniu wpływu x_1 i x_2.

---

1 Resztowa suma kwadratów będzie omówiona poniżej

Po wprowadzeniu ostatniej zmiennej, w ostatnim $(p + 1)$-szym wierszu mamy wartości (ze znakiem przeciwnym) wszystkich cząstkowych współczynników regresji oraz resztowej sumy kwadratów, natomiast w wierszach 1, 2, ..., p i kolumnach 1, 2, ..., p — a więc tam, gdzie na początku występowała macierz S — mamy elementy macierzy odwrotnej S^{-1} (ze znakami przeciwnymi). Macierz uzyskaną z U po wprowadzeniu wszystkich zmiennych niezależnych można więc przedstawić schematem

$$\begin{bmatrix} -S^{-1} & -b \\ -b^T & S_e \end{bmatrix}$$

gdzie b oznacza wektor współczynników regresji, a S_e resztową sumę kwadratów.

Zaletą metody Gaussa-Jordana jest więc to, że niezależnie od uzyskania końcowych wyników, pozwala na otrzymanie „po drodze" wyników dla mniejszej liczby zmiennych. Metodę tę można również zastosować, gdy z równania regresji wielokrotnej chce się wyeliminować jakąś zmienną. Problem taki występuje na przykład wtedy, gdy ze zbioru zmiennych, które mogą wpływać na zmienną zależną y, należy wybrać zmienne o największym wpływie. Wtedy na początku określa się równanie regresji, uwzględniające wszystkie zmienne niezależne, a następnie eliminuje się je kolejno, badając, jaki ma to wpływ na dokładność równania określoną na przykład przez resztową sumę kwadratów.

UWAGA: W praktyce często okazuje się, że z uwagi na skorelowanie zmiennych i różnice w ich wartościach o kilka a nawet kilkanaście rzędów wielkości, macierz S jest źle uwarunkowana (tzn. niemal osobliwa). Ponieważ z reguły obliczenia regresji wielokrotnej powierza się komputerowi, zatem należy upewnić się, czy zaimplementowane procedury obliczeniowe posiadają wystarczającą precyzję obliczeń. Najprostszym testem są kontrolne wydruki iloczynu macierzy S i S^{-1}, który powinien być równy macierzy jednostkowej ($SS^{-1} = I$). W przeciwnym przypadku może okazać się, że macierz odwrotna do S nie stanowi dostatecznie dobrego przybliżenia i otrzymamy błędne wyniki.

12.3 Błędy standardowe predykcji i współczynników regresji

Rozwiązanie układu równań normalnych dostarcza ocen cząstkowych współczynników regresji i jeśli na tej podstawie napiszemy równanie regresji, to możemy go użyć do przewidywania wartości y na podstawie zespołu zmiennych niezależnych $x_1, x_2, ..., x_p$. Jednakże, jeżeli rozbieżność między wartościami obserwowanymi y_i a przewidywanymi $\hat{y}_i$ będzie duża, stosowanie równania regresji nie będzie miało sensu. Konieczne jest zatem

wprowadzenie odpowiednich mierników, określających globalną dokładność predykcji oraz szacujących istotność cząstkowych współczynników regresji.

Dokładność danego równania regresji wielokrotnej możemy określić za pomocą „odległości" między y_i oraz $\hat{y}_i$, czyli za pomocą różnic $(y_i - \hat{y}_i)$, a ściślej — kwadratów tych różnic.

Sumę

$$S_e = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \ , \tag{12.8}$$

gdzie n oznacza liczebność próby, nazywa się sumą kwadratów odchyleń od regresji lub resztową sumą kwadratów. Natomiast wielkość

$$s_e = \sqrt{\frac{S_e}{n-p}} \tag{12.9}$$

nazywa się błędem standardowym predykcji; jest ona miarą dokładności przewidywania na podstawie równania

$$y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p \ .$$

Im wartość s_e jest mniejsza, tym mniejsze różnice występują między przewidywanymi na podstawie równania regresji i obserwowanymi wartościami cechy Y. Można również wykazać, że s_e^2 jest estymatorem wariancji zmiennej losowej e_i występującej w równaniu (12.1).

Określanie błędu standardowego predykcji na podstawie wartości S_e danej wzorem (12.8), jest o tyle niewygodne, że najpierw należy uzyskać oceny cząstkowych współczynników regresji b_0, b_1, ..., b_p, następnie wyliczyć z równania regresji wartości przewidywane $\hat{y}_i$ i dopiero wtedy zastosować wzór (12.8). Prościej jest obliczyć S_e ze wzoru:

$$S_e = \sum_{i=1}^{n} y_i^2 - R\,(b_0, b_1, ..., b_p) \ , \tag{12.10}$$

w którym pierwszy składnik oznacza sumę kwadratów obserwacji cechy Y (czyli zmiennej zależnej), natomiast drugi składnik $R\,(b_0, b_1, ..., b_p)$ to tzw. redukcja sumy kwadratów w wyniku dopasowania wymienionych w nawiasach stałych.

Wielkoć redukcji oblicza się ze wzoru:

$$R\,(b_0, b_1, ..., b_p) = \boldsymbol{b}^T \boldsymbol{w} = b_0 \sum_i y_i + b_1 \sum_i x_{1i}\, y_i + \ldots + b_p \sum_i x_{pi}\, y_i \tag{12.11}$$

a więc jest ona równa sumie iloczynów uzyskanych stałych b_0, b_1, ..., b_p przez wyrazy wolne równań normalnych.

Gdy równania normalne są postaci (12.6), wtedy

$$S_e = S_{yy} - R\,(b_1, ..., b_p) \tag{12.12}$$

gdzie

$$S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

jest sumą kwadratów odchyleń wartości zmiennej zależnej y od jej wartości średniej $\bar{y}$, natomiast

$$R\,(b_1, ..., b_p) = \boldsymbol{b}^T \boldsymbol{s}_y = b_1 S_{1y} + \ldots + b_p S_{py} \tag{12.13}$$

Dla równań postaci (12.9):

$$S_e = (n - p - 1)\,(s_y^2 - \boldsymbol{b}^T \boldsymbol{c}_y) = (n - p - 1)\,(s_y^2 - b_1 s_{1y} - b_2 s_{2y} - \ldots - b_p s_{py})\,, \tag{12.14}$$

gdzie

$$s_y^2 = \frac{S_{yy}}{n - 1}$$

Aby uzyskać współczynniki regresji oraz błąd standardowy predykcji, wystarczy rozwiązać układ równań normalnych, chcąc natomiast znać również błędy standardowe współczynników regresji, należy rozwiązać równanie macierzowe w celu uzyskania macierzy odwrotnej $\boldsymbol{A}^{-1}$ (bądź $\boldsymbol{S}^{-1}$) układu równań normalnych. Oznaczmy przez a^{ij} element z i-tego wiersza i j-tej kolumny macierzy odwrotnej $\boldsymbol{A}^{-1}$. Błąd standardowy cząstkowego współczynnika regresji b_j wyraża się wzorem

$$s_{b_j} = \sqrt{a^{jj}\, s_e^2}$$

i jest obojętne, czy równania normalne są postaci (12.3) czy te (12.6), ponieważ odpowiednie elementy diagonalne macierzy $\boldsymbol{A}^{-1}$ oraz $\boldsymbol{S}^{-1}$ są identyczne.

Estymator cząstkowego współczynnika regresji ma rozkład normalny z wartością oczekiwaną β_j (cząstkowy współczynnik regresji w populacji) oraz wariancją $\sigma_{b_j}^2$; estymatorem tej wariancji jest $s_{b_j}^2$.

Do weryfikacji hipotezy, że współczynnik β_j w populacji jest równy określonej wartości β_{j0}, możemy więc zastosować test t Studenta[2]. Podobnie jak w przypadku współczynnika regresji prostej, statystyka

$$t = \frac{|b_j - \beta_{j0}|}{s_{b_j}}$$

ma rozkład t Studenta o $\nu = n - p - 1$ stopniach swobody.

W szczególności, gdy $\beta_{j0} = 0$, a więc gdy sprawdzamy hipotezę o braku bezpośredniego związku między y oraz x_j mamy:

$$t = \frac{|b_j|}{s_{b_j}} \tag{12.15}$$

Hipotezę $H_0 : b_j = 0$ odrzucamy, gdy wyliczona wartość

$$t > t_{\alpha,\nu} ,$$

gdzie $t_{\alpha,\nu}$ jest wartością krytyczną rozkładu t dla poziomu istotności α i $\nu = n - p - 1$ stopni swobody.

Zwykle zmienne, dla których wyliczona wartość statystyki $t < t_{\alpha,\nu}$ odrzucamy z modelu regresji jako nieistotne i powtarzamy analizę, wyliczając ponownie wektor $\boldsymbol{b}$ dla $p' = p - 1$. Jeśli równocześnie kilka różnych zmiennych objaśniających okaże się nieistotnymi, to odrzucamy tylko jedną z nich o najmniejszej wartości funkcji testowej t i powtarzamy analizę. Należy bowiem pamiętać, że test szczegółowy pozwala na sprawdzenie istotności wprowadzenia danej zmiennej do modelu, przy założeniu, że pozostałe są tam uwzględnione. Stąd też rola innych zmiennych może się znacznie zmienić, gdy usuwamy którąkolwiek ze zmiennych.

Przykład 2.

Sprawdzimy istotność współczynników regresji uzyskanych w przykładzie 1. Do testowania hipotez $H_0 : \beta_1 = 0$, $H_0 : \beta_2 = 0$ oraz $H_0 : \beta_3 = 0$ zastosujemy wzór (12.15),

---

2 Hipotezy szczegółowe dotyczące poszczególnych cząstkowych współczynników regresji testujemy dopiero po odrzuceniu hipotezy globalnej $H_0 : \boldsymbol{b} = \mathbf{0}$; sposób testowania tej hipotezy opiszemy poniżej, omawiając analizę wariancji w regresji.

musimy więc wcześniej znać wartości błędów standardowych. Błędy te oraz odpowiadające im wartości statystyki t są następujące:

Współczynnik	Wartość	Błąd standard.	t
β_1	–0,079	0,2394	0,33
β_2	–0,397	0,4377	0,91
β_3	1,380	0,3740	3,69

Wartocią krytyczną dla $\nu = n - p - 1 = 16 - 3 - 1 = 12$ stopni swobody i dla poziomu istotności $\alpha = 5\%$ jest

$$t_{0,05;12} = 2,179$$

Możemy zatem odrzucić hipotezę o nieistotności współczynnika regresji b_3, nie ma natomiast podstaw do odrzucenia hipotez

$$H_0 : \beta_1 = 0 \quad \text{oraz} \quad H_0 : \beta_2 = 0 \ .$$

Ponieważ najmniejszą wartość statystyki t otrzymaliśmy dla współczynnika b_1, zatem odrzucamy zmienną $x1$ i powtarzamy analizę regresji. W jej wyniku otrzymujemy równanie:

$x4 = -3,208 - 0,465x2 + 1,385x3$

oraz błędy standardowe współczynników regresji[3] dane poniższą tabelą:

Współczynnik	Wartość	Błąd standard.	t
β_1	–0,465	0,3735	1,25
β_2	1,385	0,3606	3,84

Ponieważ wartość krytyczna testu t wynosi tym razem

$$t_{0,05;13} = 2,160$$

zatem okazuje się, że udział zmiennej $x2$ jest nieistotny i po ponownym przeprowadzeniu analizy pozostaje nam ostateczne równanie regresji postaci:

---

3 Uwaga: β_1 odpowiada teraz zmiennej $x2$, a β_2 — $x3$!

$x4 = -13{,}363 + 1{,}004x3$

12.4 Współczynnik korelacji wielokrotnej

Do badania jakości uzyskanego równania regresji wielokrotnej można podejść także w nieco inny sposób. Na podstawie obserwacji cechy Y w próbie (y_i) możemy obliczyć łączną sumę kwadratów S_{yy}, a w wyniku dopasowania równania regresji wielokrotnej możemy obliczyć resztową sumę kwadratów S_e. Różnica między nimi, mierząca wielkość redukcji sumy kwadratów w wyniku dopasowania (umożliwiająca obliczenie wartości przewidywanych $\bar{y}_i$), informuje o dokładności przewidywania. A zatem

$$\begin{pmatrix}\text{suma kwadratów}\\ \text{poza średnicą}\\ S_{yy}\end{pmatrix} = \begin{pmatrix}\text{suma kwadratów}\\ \text{w regresji}\\ \text{redukcja}\end{pmatrix} + \begin{pmatrix}\text{suma kwadratów}\\ \text{poza regresją}\\ S_e\end{pmatrix}$$

Jak widać z powyższego schematu, stwierdzenie „dobroci" regresji (czyli stwierdzenie, jak dalece dany model odzwierciedla rzeczywisty, uzyskany empirycznie układ punktów) sprowadza się do stwierdzenia, jak duża część sumy kwadratów w regresji (redukcji) pokrywa się z łączną sumą kwadratów (tzn. poza średnią). Użyteczną miarą tak rozumianej „dobroci" modelu regresji jest wielkość:

$$R = R_{y \cdot 12 \ldots p} = \sqrt{\frac{S_{yy} - S_e}{S_{yy}}} \tag{12.16}$$

nosząca nazwę współczynnika korelacji wielokrotnej. Jest to współczynnik korelacji wartości obserwowanych (y_i), oraz wartości obliczonych (przewidywanych $(\bar{y}_i)$, z równania regresji wielokrotnej, w którym uwzględniono zmienne niezależne $x_1, x_2, \ldots, x_p$.
Można więc inaczej zapisać, że

$$R = r_{y\bar{y}}$$

Jak wynika ze wzoru (12.16) współczynnik korelacji wielokrotnej R jest liczbą nieujemną przyjmującą wartości z przedziału <0, 1>.
Osiąga on wartość 0, gdy $S_{yy} = S_e$, a zatem gdy w wyniku dopasowania równania regresji nie nastąpiła redukcja sumy kwadratów; natomiast wartość 1, gdy $S_e = 0$, a więc gdy cała suma kwadratów została zredukowana.

Bezpośrednio ze wzoru (12.12) uzyskujemy jeszcze jeden wzór określający współczynnik korelacji wielokrotnej:

$$R^2 = \frac{R(b_1, b_2, ..., b_p)}{S_{yy}} \tag{12.17}$$

Kwadrat współczynnika korelacji wielokrotnej nazywa się współczynnikiem determinacji. Określa on, jaką część sumy kwadratów zmiennej y można wyeliminować, używając równania regresji wielokrotnej[4]. A zatem R^2 jest bezpośrednią miarą wielkości wariancji wyjaśnianej w modelu.

Jeżeli prawdziwa jest hipoteza

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0 \tag{12.18}$$

to statystyka

$$F = \frac{R^2}{1 - R^2} \cdot \frac{n - p - 1}{p} \tag{12.19}$$

ma rozkład F o $\nu_1 = p$ oraz $\nu_2 = (n - p - 1)$ stopniach swobody. Statystyki tej można użyć do testowania hipotezy (12.18), która jest równoważna z hipotezą

$$H_0 : R = 0$$

Aby test F mógł być stosowany, zmienna zależna w równaniu regresji musi mieć rozkład normalny, nie są natomiast potrzebne żadne założenia o rozkładzie zmiennych niezależnych. Uwzględniając w (12.19) wzór definiujący współczynnik korelacji wielokrotnej[5], po prostych przekształceniach otrzymujemy:

$$F = \left(\frac{S_{yy} - S_e}{p}\right) : \left(\frac{S_e}{n - p - 1}\right), ,$$

a jeśli jeszcze wykorzystamy związek (12.12), uzyskamy kolejną wersję wzoru na statystykę F:

$$F = \left(\frac{R(b_1, b_2, ..., b_p)}{p}\right) : \left(\frac{S_e}{n - p - 1}\right) \tag{12.20}$$

---

4 Oczywiście eliminacja będzie całkowita, jeżeli wartości przewidywane $\hat{y}_i$ będą identyczne z wartościami obserwowanymi y_i

5 czyli wzór (12.16)

Przy interpretacji współczynnika korelacji wielokrotnej należy pamiętać o wielkości próby. Czynnik ten odgrywa podstawową rolę w tzw. zjawisku „kurczenia się prognozy". Zjawisko to polega na przeszacowaniu wartości współczynnika R w przypadku prób mało licznych. Należy zaznaczyć, że zjawisko kurczenia się prognozy nie występuje dopiero wtedy, gdy liczba elementów próby jest duża, przynajmniej trzydzieści razy większa od niż liczba zmiennych.

Niektórzy autorzy sugerują w przypadku małej próby użycie odpowiednich wzorów, pozwalających na przeliczenie uzyskanej wartości kwadratu współczynnika korelacji wielokrotnej. Przytoczymy tu jedynie podstawowy z nich, a mianowicie przeliczona wartość współczynnika korelacji wielokrotnej, oznaczana jako $\hat{R}^2$, wynosi:

$$\hat{R}^2 = 1 - \frac{n-3}{n-p-1}(1-R^2) + \frac{2}{n-p+1}(1-R^2)^2$$

Przykład 3.

W poprzednim przykładzie testowaliśmy istotność cząstkowych współczynników regresji. Obecnie zajmiemy się testem istotności współczynnika korelacji wielokrotnej. Resztowa suma kwadratów wynosi S_e = 1333,11, natomiast łączna suma kwadratów jest równa S_{yy} = 4395,48. Otrzymujemy zatem (zgodnie z wzorami (12.16) i (12.17)):

$$R^2 = 0{,}6967, \qquad R = 0{,}837$$

Testujemy hipotezę $H_0 : R = 0$ równoważną hipotezie (12.18). Wyliczona zgodnie ze wzorem (12.19) statystyka testowa F ma wartość F = 9,188. Ponieważ wartością krytyczną dla $\nu_1 = 3$ i $\nu_2 = 12$ stopni swobody oraz dla poziomu istotności 5% jest $F_{0,05;2;12}$ = 3,490, zatem

$$F > F_{0,05;2;12}$$

i hipotezę zerową odrzucamy. Oznacza to jednocześnie odrzucenie hipotezy $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$. Wynik ten **nie jest sprzeczny** z rezultatami poprzedniego przykładu. Wykazaliśmy bowiem teraz, że nie można przyjąć, iż wszystkie trzy współczynniki β są jednocześnie równe zeru — nie przesądza to jednak o nieistotności niektórych z nich.

12.5 Analiza wariancji w regresji

Hipotezy formułowane w poprzednich punktach dotyczyły cząstkowych współczynników regresji oraz współczynnika korelacji wielokrotnej, jednak ich weryfikacja odbywała

się na podstawie sum kwadratów odchyleń. Całkowita suma kwadratów dzieli się, jak już to wiemy, na dwa składniki:

$$S_{yy} = R\,(b_1, ..., b_p) + S_e$$

Sumie

$$S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

odpowiada $(n - 1)$ stopni swobody, redukcji $R\,(b_1, ..., b_p)$ odpowiada p stopni swobody, wreszcie sumie resztowej S_e odpowiada $(n - p - 1)$ stopni swobody. Dzieląc każdą z sum kwadratów przez liczbę stopni swobody otrzymujemy wariancje:

$$s_y^2 = \frac{S_{yy}}{n-1}, \qquad s_{\text{reg}}^2 = \frac{R\,(b_1, ..., b_p)}{p}, \qquad s_e^2 = \frac{S_e}{n-p-1}$$

Wariancje s_{reg}^2 i s_e^2 występują we wzorze (12. 20), który możemy teraz przedstawić w postaci:

$$F = \frac{s_{\text{reg}}^2}{s_e^2}$$

Uzyskany wzór i przeprowadzone postępowanie są bardzo podobne do stosowanego w analizie wariancji i dlatego wyniki analizy dotyczącej regresji na ogół przedstawia się w tabeli podobnej do tabeli analizy wariancji. Takie przedstawienie wyników nazywane jest analizą wariancji w regresji lub krócej analizą regresji.

Schemat analizy regresji przedstawia tabela 12.1. W łatwy sposób można się przekonać, że jest on odpowiedni również i dla regresji z jedną tylko zmienną niezależną (tzn. $p = 1$).

Mając określone równanie z p zmiennymi niezależnymi, często stawia się pytanie czy wszystkie te zmienne są w równaniu potrzebne[6], a więc czy nie można by zrezygnować z części z nich — bez istotnego zmniejszenia dokładności oceny.

---

6 Odpowiedź na to pytanie próbowaliśmy już znaleźć (w nieco inny sposób) w przykładzie 2.

Tabela 12.1

Schemat analizy regresji przy hipotezie $H_0 : \beta_1 = \beta_2 = \ldots = \beta_p$

Zmienność	Liczba stopni swobody	Suma kwadratów	Średni kwadrat	F
Regresja	p	$S_{reg} = R(b_1, b_2, \ldots, b_p) = R^2 S_{yy}$	$s^2_{reg} = \frac{S_{reg}}{p}$	$\frac{s^2_{reg}}{s^2_e} = \frac{R^2}{1 - R^2}$
Odchylenie od regresji (błąd)	$n - p - 1$	$S_e = (1 - R^2) S_{yy}$	$s^2_e = \frac{S_e}{n - p - 1}$	
Całkowita	$n - 1$	S_{yy}		

Podzielmy zatem zbiór p zmiennych $x_1, x_2, \ldots, x_p$ na dwa podzbiory: w pierwszym umieśćmy k zmiennych, których uwzględnienie w równaniu uznajemy za niezbędne (są to zmienne $x_1, x_2, \ldots, x_k$), a w drugim zmienne, których wpływ na y może być nieistotny ($x_{k+1}, x_{k+2}, \ldots, x_p$). Badanie wpływu drugiej grupy zmiennych sprowadza się do weryfikacji hipotezy, że cząstkowe współczynniki regresji odpowiadające tym zmiennym są równe zeru, czyli

$$H_0 : \beta_{k+1} = \beta_{k+1} = \ldots = \beta_p = 0 \tag{12.21}$$

Hipoteza alternatywna stwierdza, że nie wszystkie współczynniki są równe 0. Redukcja sumy kwadratów wynikająca z ustalenia wszystkich współczynników regresji wynosi

$$R(b_1, \ldots, b_p) ,$$

natomiast redukcja wynikająca z ustalenia k pierwszych współczynników jest równa

$$R(b_1, \ldots, b_k) .$$

Różnica

$$R(b_1, \ldots, b_p) - R(b_1, \ldots, b_k)$$

oznacza tę część redukcji sumy kwadratów, która wynika z ustalenia dodatkowych współczynników b_{k+1}, b_{k+2}, ..., b_p.

Przy założeniu hipotezy zerowej postaci (12.21) statystyka:

$$F = \frac{s_w^2}{s_e^2}, \text{ gdzie } s_w^2 = \frac{R(b_1, b_2, ..., b_p) - R(b_1, ..., b_k)}{p - k}$$

ma rozkład F o $\nu_1 = (p - k)$ i $\nu_2 = (n - p - 1)$ stopniach swobody.

Tabela 12.2

Schemat analizy regresji przy hipotezie $H_0 : \beta_{k+1} = \beta_{k+2} = \ldots = \beta_p$, $\quad k < p$

Zmienność	Liczba stopni swobody	Suma kwadratów	Średni kwadrat	F
Regresja względem $b_1, b_2, ..., b_p$	p	$R_1 = R(b_1, b_2, ..., b_p)$	$s_{reg}^2 = \frac{R_1}{p}$	$F^1 = \frac{s_{reg}^2}{s_e^2}$
Regresja względem $b_1, b_2, ..., b_k$	k	$R_2 = R(b_1, b_2, ..., b_k)$		
Regresja względem $b_{k+1}, ..., b_p$	$p - k$	$R_1 - R_2$	$s_w^2 = \frac{R_1 - R_2}{p - k}$	$F^0 = \frac{s_w^2}{s_e^2}$
Błąd	$n - p - 1$	$S_e = S_{yy} - R_1$	$s_e^2 = \frac{S_e}{n - p - 1}$	
Całkowita	$n - 1$	S_{yy}		

Schemat analizy wariancji w omawianym przypadku podaje tabela 12.2. W tabeli tej zaznaczono dwa wyrażenia F. Pierwsze z nich, F^1, weryfikuje hipotezę (12.18), dotyczącą całego równania regresji (tzn. wszystkich zmiennych), drugie zaś, F^0, weryfikuje hipotezę (12.21), dotyczącą części zmiennych.

Jeżeli hipoteza (12.21) nie zostanie odrzucona, możemy z równania regresji usunąć zmienne x_{k+1}, x_{k+2}, ..., x_p, należy jednak pamiętać, że nieodrzucenie hipotezy H_0 nie jest

równoznaczne z jej prawdziwością. Dlatego też usunięcie nawet i nieistotnych zmiennych zmniejszy nieco dokładność równania.

Moc testu F zależy od liczebności próby, gdy więc próba jest niezbyt liczna, muszą wystąpić duże różnice, aby je uznać za istotne. Powinniśmy więc dążyć do tego, aby występujące w równaniach normalnych sumy kwadratów i iloczynów były obliczone na podstawie jak największej liczby obserwacji.

Przykład 4.

Przytoczymy teraz tabelę analizy wariancji dla danych rozpatrywanych w poprzednich przykładach. Testujemy zatem hipotezę $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$.

Zmienność	Liczba stopni swobody	Suma kwadratów	Średni kwadrat	F
Regresja	3	3062,37	1020,79	9,189
Odchylenie od regr.	12	1333,11	111,092	
Całkowita	15	4395,48		

Wyliczona wartość F = 9,189 jest większa od wartości krytycznej wynoszącej $F_{0,05;3;12}$ = 3,490, zatem odrzucamy postawioną hipotezę zerową. Uzyskany wynik pokrywa się oczywiście z wynikiem otrzymanym w przykładzie 2.

12.6 Regresja krokowa

Przystępując do budowy modelu regresji wielokrotnej kierujemy się zazwyczaj dwoma sprzecznymi kryteriami:

1. Aby uzyskać równanie przydatne do celów predykcji dążymy do wprowadzenia do modelu tak wielu zmiennych niezależnych, jak to jest możliwe, gdyż im więcej jest w modelu uwzględnionych zmiennych niezależnych, tym lepiej, pełniej wyjaśniona będzie zmienna zależna.
2. Ze względu na koszty związane z uzyskaniem informacji o dużej liczbie zmiennych i czas zużyty na kolejne ich uwzględnianie w modelu chcielibyśmy, aby równanie zawierało jak najmniej zmiennych niezależnych.

Oba kryteria uwzględniane łącznie prowadzą do określenia optymalnego zestawu zmiennych niezależnych, do uzyskania optymalnego równania regresji wielokrotnej.

W zasadzie wyróżnić można trzy rodzaje procedur wprowadzania zmiennych do modelu regresji:

— procedurę wszystkich możliwych regresji,
— procedurę eliminacji *a posteriori*,
— procedurę regresji krokowej.

Procedura wszystkich możliwych regresji jest procedurą bardzo czasochłonną i uciążliwą. Jak sama nazwa wskazuje, wymaga ona przeanalizowania wszystkich możliwych równań regresji. Mając zatem p zmiennych niezależnych, musimy zbudować i przeanalizować 2^p równań (dla $p = 4$ liczba ta wynosi 16, ale już dla $p = 8$ jest to 256!). Dysponując zestawem wszystkich równań, wybieramy to równanie, które wyjaśnia najwięcej wariancji zmiennej zależnej. Kryterium wyboru stanowi tutaj wartość R^2 — wartość kwadratu współczynnika korelacji wielokrotnej. Zdarzyć się może, że kilka równań posiada tę samą lub zbliżoną wartość R^2, a wówczas wybieramy optymalne równanie kierując się dodatkowym kryterium , tzn. skorelowaniem zmiennych niezależnych między sobą oraz badając średnie kwadraty resztowe. Jeśli weźmie się pod uwagę przyrost równań wraz ze wzrostem ilości zmiennych niezależnych, to wybór równania optymalnego nie wydaje się być sprawą prostą. Praktycznie powyżej kilku zmiennych przestaje się panować nad materiałem uzyskanym z maszyny cyfrowej.

Procedura eliminacji *a posteriori* nie wymaga już analizowania takiej dużej liczby równań jak procedura wszystkich możliwych regresji. Zasadnicze kroki tej procedury są następujące:

1. Obliczamy równanie regresji zawierające wszystkie zmienne niezależne.
2. Przeprowadzamy obliczenia częściowego testu F (w sposób omówiony w poprzednim podrozdziale) dla każdej sprawdzanej zmiennej niezależnej, jak gdyby to była ostatnia zmienna niezależna wchodząca do równania regresji.
3. Porównujemy najmniejszą wartość F z częściowego testu F, powiedzmy $F_{\min}$, z wartością dla wstępnie obranego poziomu istotności, powiedzmy F_α,
   — jeżeli $F_{\min} < F_\alpha$, usuwamy z rozważań zmienną niezależną X, z której wynikało $F_{\min}$, ponownie obliczamy równanie regresji z pozostałymi zmiennymi niezależnymi i powracamy do kroku 2,
   — Jeżeli $F_{\min} > F_\alpha$, przyjmujemy równanie regresji zgodnie z obliczeniem.

Po dokonaniu eliminacji w równaniu zostają tylko te zmienne niezależne, które w sposób istotny wyjaśniają wariancję zmiennej zależnej. Procedura eliminacji *a posteriori* jest mniej czasochłonna od procedury wszystkich możliwych regresji. Może być ona wykorzystywana z powodzeniem przy większej liczbie zmiennych. Jest to na ogół bardzo korzystna procedura, w szczególności jeśli chcemy widzieć wszystkie zmienne niezależne w jednym równaniu, aby „czegoś nie opuścić". Jednakże, jeżeli dane wejściowe dają macierz układu równań normalnych, która jest źle uwarunkowana, wtedy ta procedura może prowadzić do niedorzeczności wskutek błędów zaokrąglania. Problem ten nie istnieje jeśli korzystamy z profesjonalnego oprogramowania.

Krokowe procedury wprowadzania zmiennych niezależnych do liniowego modelu regresji są jak gdyby odwróceniem postępowania procedurze eliminacji *a posteriori*. W procedurach krokowych model tworzy się poprzez wprowadzenie kolejno (w kolejnych krokach) poszczególnych zmiennych. Kryterium wprowadzenia stanowi tutaj wielkość wariancji wyjaśnianej przez daną zmienną. W pierwszej kolejności do modelu wprowadzana jest zmienna, która wyjaśnia najwięcej wariancji zmiennej zależnej. Następnie do modelu wprowadzane są zmienne wyjaśniające odpowiednio największe części pozostałej, nie wyjaśnionej jeszcze, wariancji. Na uwagę zasługują dwa rodzaje procedur krokowych.

Pierwsza z nich, procedura selekcji *a priori*, polega na sukcesywnym wprowadzaniu zmiennych do modelu w omówiony wyżej sposób. W efekcie otrzymujemy równanie regresji, zawierające zmienne uporządkowane zgodnie z ich udziałem w wyjaśnianiu zmienności zmiennej zależnej. Opierając się na teście F częściowym sprawdzamy, do którego kroku wprowadzane zmienne wyjaśniają istotne części wariancji zmiennej zależnej. Innymi słowy sprawdzamy, kiedy wartość testu F związana z ostatnio wprowadzoną zmienną przestaje być istotna. W tym momencie kończymy proces, przyjmując otrzymane równanie jako ostateczny model. Metoda selekcji *a priori* jest często stosowana w praktyce, a jej niewielką wadą jest to, że nie pokazuje wpływu, jaki może mieć wprowadzenie nowej zmiennej niezależnej na znaczenie zmiennej wprowadzonej na wcześniejszym etapie. Trudność tę można pokonać stosując opisaną poniżej procedurę pełnej regresji krokowej.

Druga odmiana procedury krokowej, tzw. procedura pełnej regresji krokowej, jest procedurą minimalizującą w modelu liczbę zmiennych. W procedurze tej, podobnie jak w poprzedniej, zmienne wprowadzane są do modelu kolejno — w zależności od wielkości wariancji, jaką wyjaśniają. Ulepszenia polegają na powtórnym badaniu na każdym etapie regresji zmiennych niezależnych, wprowadzonych do modelu w poprzednich etapach. Zmienna niezależna, która mogła być najlepszą pojedynczą zmienną do wprowadzenia w poprzedzającym etapie, może w etapie późniejszym być zbyteczna ze względu na swoją zależność od innych zmiennych niezależnych występujących teraz w regresji. Dla sprawdzenia tego ocenia się i porównuje kryterium wartości F dla każdej zmiennej niezależnej w regresji na każdym etapie obliczeń z wstępnie wybraną wartością krytyczną odpowiedniego rozkładu F. Daje to informację o udziale każdej zmiennej niezależnej, tak jak gdyby to była najwcześniej wprowadzona zmienna niezależna, bez względu na rzeczywisty moment wprowadzenia jej do modelu. Każda zmienna niezależna, której udział jest nieistotny, jest usuwana z modelu. Proces ten trwa tak długo, aż żadna ze zmiennych niezależnych nie będzie mogła być dopuszczona do równania i żadna odrzucona. Jak więc widać, pełna procedura krokowa jest bardzo elegancka. Jest ona nastawiona na zbudowanie takiego modelu regresji, w którym zawarte są jedynie niezbędne zmienne niezależne z punktu widzenia wyjaśniania wariancji zmiennej zależnej. Procedura pełnej regresji krokowej wydaje się być zdecydowanie najlepsza z omówionych procedur wyboru zmiennych niezależnych. Warto jednak pamiętać, że może być ona niewłaściwie wyko-

rzystana przez statystyka „amatora", gdyż bardzo łatwo jest polegać zbyt dosłownie na wyborze automatycznym dokonanym przez komputer. W praktycznych zastosowaniach procedura pełnej regresji krokowej wzbogacona jest na ogół o pewne opcje, jak np. możliwość wskazania zmiennych niezależnych, które muszą znaleźć się w modelu, niezależnie od wartości testu F.

Przykład 5.

Zastosowanie procedury pełnej regresji krokowej do danych z przykładu 1 daje wynik zgodny z oczekiwaniami (por. przykład 2 i 3), tzn. w pierwszym kroku do równania regresji wprowadzana jest zmienna $x3$, a następnie nie ma już podstaw do wprowadzania kolejnych zmiennych. Końcowa postać równania regresji jest zatem następująca:

$$x4 = -13{,}363 + 1{,}004x3.$$

12.7 Standaryzowane cząstkowe współczynniki regresji

Współczynniki b_j równania regresji wielokrotnej (5) wiążą się ściśle z jednostkami pomiaru poszczególnych cech. Jeżeli chcemy porównać te współczynniki między sobą, w celu ustalenia ważności poszczególnych cech występujących w równaniu regresji, musimy dokonać standaryzacji zmiennych y i x_i za pomocą przekształceń

$$y' = \frac{y - \bar{y}}{s_y}, \qquad x_i' = \frac{x_i - \bar{x}_i}{s_i} \qquad (i = 1, \ldots, p)$$

gdzie s_y oznacza standardowe odchylenie zmiennej zależnej y, natomiast s_i to standardowe odchylenie zmiennej niezależnej x_i. Wtedy wariancje standaryzowanych zmiennych są równe jednościom, a równanie regresji można przedstawić w postaci

$$y' = b_1' x_1' + b_2' x_2' + \ldots + b_p' x_p'$$

lub

$$\frac{y - \bar{y}}{s_y} = b_1' \frac{x_1 - \bar{x}_1}{s_1} + b_2' \frac{x_2 - \bar{x}_2}{s_2} + \ldots + b_p' \frac{x_p - \bar{x}_p}{s_p} \qquad (12.22)$$

Współczynniki b_i' nazywa się standaryzowanymi cząstkowymi współczynnikami regresji. Są one wyrażone w jednostkach stosunku s_y / s_i a zatem są liczbami niemianowanymi, co pozwala na porównanie ich między sobą. Znając współczynniki b_i' łatwo można przejść

do „zwykłych", czyli niestandaryzowanych współczynników regresji (tzn. występujących w równaniu regresji wielokrotnej — por. zależność (5)):

$$b_i = b_i' \frac{s_y}{s_i}, \tag{12.23}$$

ponieważ, mnożąc obie strony (12.22) przez s_y, otrzymujemy

$$y - \bar{y} = b_1' \frac{s_y}{s_1}(x_1 - \bar{x}_1) + b_2' \frac{s_y}{s_2}(x_2 - \bar{x}_2) + \ldots + b_p' \frac{s_y}{s_p}(x_p - \bar{x}_p)$$

Równania normalne do oceny współczynników b_i' będą miały teraz postać

$$\begin{bmatrix} 1 & r_{12} & \ldots & r_{1p} \\ r_{12} & 1 & \ldots & r_{2p} \\ \ldots & \ldots & \ldots & \ldots \\ r_{1p} & r_{2p} & \ldots & 1 \end{bmatrix} \begin{bmatrix} b_1' \\ b_2' \\ . \\ . \\ b_p' \end{bmatrix} = \begin{bmatrix} r_{1y} \\ r_{2y} \\ . \\ . \\ r_{py} \end{bmatrix} \tag{12.24}$$

gdzie r_{ij} oznacza współczynnik korelacji x_i oraz x_j, natomiast r_{iy} oznacza współczynnik korelacji zmiennej zależnej y i niezależnej x_i.

Ocena współczynników regresji na podstawie układu równań (12.24) jest wygodna z tego względu, że występujące w tym układzie wielkości r_{ij} oraz r_{iy} są liczbami podobnego rzędu wielkości (od –1 do 1), stąd też wszystkie działania arytmetyczne mogą być wykonane z tą samą dokładnością. Po uzyskaniu współczynników b_i' wartości niestandaryzowane wyliczy się ze wzoru (12.23), a współczynnik korelacji wielokrotnej — ze wzoru

$$R^2 = b_1' r_{1y} + b_2' r_{2y} + \ldots + b_p' r_{py} \tag{12.25}$$

Współczynnik korelacji wielokrotnej przyjmuje oczywiście taką samą wartość dla zmiennych standaryzowanych jak i niestandaryzowanych. Jak wynika z tabeli 11.1, do oceny istotności regresji wystarczy znać wartość R^2, ponieważ

$$F = \frac{R^2}{1 - R^2} .$$

Dodatkowo zachodzi

$$R(b_1', b_2', \ldots, b_p') = R^2$$

Jeżeli przez S_e' oznaczymy resztową sumę kwadratów dla zmiennej standaryzowanej y', to

$$S_e' = S_{y'y'} - R\,(b_1', b_2', \ldots, b_p') = (n - p - 1)\,(1 - R^2)$$

a zatem błąd standardowy predykcji zmiennej standaryzowanej wynosi

$$s_e' = \sqrt{\frac{S_e'}{n - p - 1}} = \sqrt{1 - R^2}$$

12.8 Współczynnik korelacji cząstkowej

Rozpatrzmy trzy cechy oznaczone x_1, x_2, x_3. Załóżmy, że współczynnik korelacji r_{12} między x_1 a x_2 jest wysoki, co jednak wynika nie tyle z faktycznej zależności między x_1 a x_2, lecz z powiązania obu tych cech z x_3. Gdy utworzymy równania regresji określające x_1 i x_2 na podstawie x_3, tzn.:

$$\hat{x}_1 = \bar{x}_1 + b_1\,(x_3 - \bar{x}_3)$$

$$\hat{x}_2 = \bar{x}_2 + b_2\,(x_3 - \bar{x}_3)$$

wówczas zmienne

$$x_{1.3} = x_1 - \hat{x}_1 \qquad \text{oraz} \qquad x_{2.3} = x_2 - \hat{x}_2$$

obrazują, jaka część wartości x_1 lub x_2 nie została wyeliminowana w wyniku ustalenia zmiennej x_3. Współczynnik korelacji $x_{1.3}$ oraz $x_{2.3}$ nazywa się współczynnikiem korelacji cząstkowej zmiennych x_1 i x_2, przy ustalonej zmiennej x_3 i oznacza symbolem $r_{12.3}$. Oblicza się go ze wzoru:

$$r_{12.3} = \frac{r_{12} - r_{13}\,r_{23}}{\sqrt{(1 - r_{13}^2)\,(1 - r_{23}^2)}} \tag{12.26}$$

gdzie r_{ij} oznacza zwykły współczynnik korelacji zmiennych x_i oraz x_j. Ogólnie, współczynnik korelacji cząstkowej dwóch zmiennych przy ustalonej trzeciej określamy wzorem

$$r_{ij.k} = \frac{r_{ij} - r_{ik}\,r_{jk}}{\sqrt{(1 - r_{ik}^2)\,(1 - r_{jk}^2)}} \tag{12.27}$$

Gdy rozpatrujemy cztery cechy, wtedy współczynnik korelacji cząstkowej dwóch z nich (i, j) przy ustalonych dwóch pozostałych (k, l) wyraża się wzorem

$$r_{ij.kl} = \frac{r_{ij.k} - r_{il.k} r_{jl.k}}{\sqrt{(1 - r_{il.k}^2)(1 - r_{jl.k}^2)}} \tag{12.28}$$

a zatem współczynnik korelacji cząstkowej z dwiema ustalonymi zmiennymi można obliczyć za pośrednictwem współczynników korelacji cząstkowej z jedną ustaloną zmienną. Zamiast k oraz l można wstawić dowolną grupę wskaźników i w ten sposób stworzyć możliwość obliczenia współczynników korelacji cząstkowej dla większej liczby ustalonych zmiennych niezależnych.

Ogólnie, jeśli $\boldsymbol{R}$ oznacza macierz korelacji między zmiennymi, to współczynnik korelacji cząstkowej $r_{12.3...k}$ między zmiennymi x_1 i x_2 przy ustalonych pozostałych zmiennych x_3, ..., x_k wyraża się wzorem

$$r_{12.3...k} = \frac{|\boldsymbol{R}_{12}|}{\sqrt{|\boldsymbol{R}_{11}| \cdot |\boldsymbol{R}_{22}|}} , \tag{12.29}$$

gdzie $|\boldsymbol{R}_{ij}|$ jest dopełnieniem algebraicznym elementu r_{ij} macierzy $\boldsymbol{R}$.

Gdy rozpatrujemy 3 zmienne, macierz

$$\boldsymbol{R} = \begin{bmatrix} 1 & r_{12} & r_{13} \\ r_{12} & 1 & r_{23} \\ r_{13} & r_{23} & 1 \end{bmatrix}$$

stąd

$$|\boldsymbol{R}_{11}| = (-1)^{1+1} \begin{vmatrix} 1 & r_{23} \\ r_{23} & 1 \end{vmatrix} = 1 - r_{23}^2 ,$$

$$|\boldsymbol{R}_{22}| = (-1)^{2+2} \begin{vmatrix} 1 & r_{13} \\ r_{13} & 1 \end{vmatrix} = 1 - r_{13}^2 ,$$

$$|\boldsymbol{R}_{12}| = (-1)^{1+2} \begin{vmatrix} r_{12} & r_{23} \\ r_{13} & 1 \end{vmatrix} = -(r_{12} - r_{13} r_{23}) .$$

Podstawiając uzyskane dopełnienia do wzoru (12.29) otrzymujemy wzór (12.26). Aby uzyskać wzór (12.27) należy określić dopełnienia macierzy $\boldsymbol{R}$.

Istotność współczynnika korelacji cząstkowej można badać za pomocą testu t Studenta, na podobnych zasadach jak w przypadku zwykłego współczynnika korelacji, jedynie ze zmianą dotyczącą liczby stopni swobody. Przy założeniu hipotezy

$$H_0 : r_{12.3...k} = 0$$

statystyka

$$t = \frac{r_{12.3...k}}{\sqrt{1 - r^2_{12.3...k}}} \sqrt{n - k}$$

ma rozkład t Studenta o $\nu = (n - k)$ stopniach swobody.

13. REGRESJA KRZYWOLINIOWA

Omawiane dotychczas modele regresji miały wszystkie postać zależności liniowej między zmienną zależną a jedną lub wielu zmiennymi niezależnymi, tzn. postać funkcji pierwszego stopnia zmiennej lub zmiennych niezależnych. Ta dogodna zależność nie zawsze jednak daje zadowalającą korelację — pomocne jest wówczas zastosowanie regresji krzywoliniowej.

Modele nieliniowe można podzielić na dwa typy, które będą nazywane modelami sprowadzalnymi do liniowych i niesprowadzalnymi do liniowych. Jeśli model jest sprowadzalny do liniowego można go przedstawić przy użyciu odpowiednich przekształceń zmiennych w postaci standardowej modelu liniowego

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p \ . \tag{13.1}$$

Przykładowo, stosowany najczęściej model wielomianowy

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_p x^p \tag{13.2}$$

bazuje na podstawieniu za kolejne zmienne w równaniu (13.1) kolejnych potęg zmiennej x, tzn.

$$x_1 = x\,, \qquad x_2 = x^2\,, \qquad x_3 = x^3\,, \qquad x_4 = x^4\,, \ldots$$

Zastosowanie wyników teoretycznych z rozdziału dotyczącego regresji wielokrotnej jest wówczas bezpośrednie.

Jeżeli modelu nieliniowego nie można przedstawić w postaci (13.1), to jest on modelem niesprowadzalnym do liniowego (tzn. bezwarunkowo nieliniowym). Sięgać trzeba wówczas do metod optymalizacji nieliniowej[1].

---

1 Czytelnikom chcącym zapoznać się bliżej z tą tematyką polecamy książki:
Zieliński R.: Wybrane zagadnienia optymalizacji statystycznej, PWN, Warszawa, 1974
Zieliński R.: Stochastyczne algorytmy optymalizacji, IMPAN, Warszawa, 1980

Zajmiemy się wyłącznie skrótowym omówieniem modeli sprowadzalnych do liniowych (linearyzowalnych), gdyż są one najczęściej stosowanymi w biometrii metodami korelacji krzywoliniowej.

Jak już wspomnieliśmy, modele linearyzowalne są to modele dające się sprowadzić do modelu liniowego przez odpowiednią transformację zmiennych. Takimi modelami często stosowanymi w praktyce są:

model potęgowy

$$y = ax^b \qquad \text{dla} \quad x > 0, \quad y > 0, \quad a > 0 \ ,$$

linearyzowalny po transformacji logarytmicznej zmiennych

$$\ln y = \ln a + b \ln x$$

i podstawieniu

$$z = \ln y, \qquad u = \ln x \, ;$$

model wykładniczy

$$y = e^{a + bx} \ ,$$

po transformacji

$$\ln y = a + bx, \qquad y > 0 \, ;$$

model złożony wykładniczy

$$y = ax^b \, e^{cx} \ ,$$

po transformacji

$$\ln y = \ln a + b \ln x + cx, \qquad x > 0, \quad y > 0, \quad a > 0 \, ;$$

model wielomianowy

$$y = b_0 + b_1 x + b_2 x^2 + \ldots + b_p x^p \ , \tag{13.3}$$

sprowadzający się do modelu liniowego regresji wielokrotnej po prostym podstawieniu za kolejne zmienne w równaniu (13.3) kolejnych potęg zmiennej x. Model wielomianowy

jest najczęściej stosowany w praktycznych zastosowaniach regresji krzywoliniowej ze względu na ogromną różnorodność kształtu krzywych wielomianowych oraz fakt pozostawania zmiennej y bez transformacji w tym modelu. Dąży się zazwyczaj do tego, aby stopień wielomianu p był jak najmniejszy. Za ograniczeniem wartości p przemawiają dwa argumenty:

— im więcej składników uwzględnia się w równaniu regresji, tym bardziej pracochłonne stają się obliczenia,

— z algebry wiadomo, że dla dowolnego zbioru punktów istnieje taka krzywa opisana równaniem postaci (13.3), która przechodzi dokładnie przez wszystkie punkty (dla k punktów będziemy mieli $p = k - 1$), jednak otrzymane równanie tak wysokiego rzędu raczej zagmatwa, niż rozjaśni obraz zależności.

Do wyznaczania parametrów funkcji regresji w przypadku modeli linearyzowalnych stosuje się, tak jak i poprzednio, estymatory metody najmniejszych kwadratów. Trzeba jednak pamiętać, że w metodzie najmniejszych kwadratów minimalizuje się odchylenia od modelu po linearyzacji[2]. Zatem odchylenia będą liczone na logarytmach w modelu potęgowym i wykładniczym, co powoduje, że dopasowanie krzywej regresji do danych empirycznych będzie lepsze w pewnym zakresie skali[3], a gdzie indziej gorsze. Dlatego też w regresji krzywoliniowej stosuje się czasami ważoną metodę najmniejszych kwadratów, gdzie wagami przy kwadratach odchyleń są odwrotności oczekiwanych wariancji odchyleń. W przypadku transformacji zmiennej zależnej Y po dopasowaniu krzywej regresji oblicza się odchylenia na wielkościach we właściwej, wyjściowej skali (po retransformacji) i wyznacza się ich średnią kwadratową według wzoru

$$s_e = \sqrt{\frac{1}{n} \sum (y_i - \hat{y}_i)^2} \tag{13.4}$$

gdzie $\hat{y}_i = m(x_i)$ jest to wartość funkcji regresji odpowiadająca obserwacji y_i, jako jedną z miar dopasowania krzywej do danych empirycznych, obok charakterystyk zwykle stosowanych w modelu liniowym. Kłopotu tego nie nastręcza regresja wielomianowa. Modele wielomianowe sprawiają jednak kłopoty numeryczne, gdyż kolejne potęgi zmiennych niezależnych i ich iloczyny są silnie skorelowane, ponadto ich wartości różnią się czasami o kilka rzędów wielkości, a kowariancje między nimi różnią się nawet o kilkanaście rzędów wielkości. Transformując ten model na model regresji wielokrotnej uzyskujemy macierz kowariancji C nie tylko o bardzo zróżnicowanych elementach co do rzędu

---

2 co oznacza, że trzeba być ostrożnym i sprawdzić, czy założenia metody najmiejszych kwadratów (niezależne błędy, $N(0, \sigma^2)$) są zachowane przy dokonywaniu przekształcenia

3 tu w pobliżu początku układu współrzędnych

wielkości, ale ponadto jej kolumny (wiersze) są niemal współliniowe, a zatem macierz C jest niemal osobliwa. Wtedy otrzymanie wystarczająco dokładnych[4] rozwiązań układu równań normalnych wymaga zastosowania odpowiednich procedur o dużej precyzji obliczeń. Pewnym praktycznym zabiegiem poprawiającym dokładność numeryczną obliczeń jest sprowadzanie poszczególnych zmiennych do postaci półlogarytmicznej: podzielenie każdej z nich przez taką potęgę 10, aby jej wartości miały jedną cyfrę przed przecinkiem, tzn. należały do przedziału <1, 10). Wtedy współczynniki regresji zwiększą się w tym samym stosunku, a wszelkie testy statystyczne nie ulegną zmianie.

Analizę regresji krzywoliniowej według modelu wielomianowego można znacznie ułatwić, jak również zmniejszyć pracochłonność obliczeń, gdy zmienna objaśniająca X jest kontrolowana i możemy dobrać jej wartości tak, aby tworzyły ciąg arytmetyczny, tzn.

$$x_{i+1} = x_i + n = x_1 + ih \tag{13.5}$$

Wówczas transformację modelu regresji wielomianowej (13.3) na model regresji wielokrotnej możemy przeprowadzić tak, aby poszczególne zmienne w modelu wielokrotnym były parami nieskorelowane (ortogonalne). Daje to w efekcie macierz układu równań normalnych C postaci diagonalnej. Wtedy cały układ równań normalnych rozpada się na p niezależnych równań liniowych z jedną niewiadomą każde, rozwiązuje się go więc natychmiast. Również obliczenie macierzy odwrotnej do C jest proste, mianowicie macierz C^{-1} ma na przekątnej odwrotności elementów leżących na przekątnej macierzy C, a poza tym zera, jest więc diagonalna. Dlatego też analiza istotności poszczególnych zmiennych w regresji wielokrotnej jest prosta — każda zmienna może być testowana niezależnie od innych, a kolejność testowania nie ma znaczenia.

Transformacji, o której mówimy (ma ona nazwę ξ', tzn. *ksi prim*), możemy dokonać posługując się układem wielomianów ortogonalnych. Wartości wielomianów ortogonalnych w punktach x_i spełniających warunek (13.5) tworzą właśnie układ nieskorelowanych danych[5].

Wielomiany te $\xi_j(x)$ do czwartego stopnia łącznie dane są wzorami:

$$\xi_1(x) = k(x)\lambda_1$$

$$\xi_2(x) = \left[k^2(x) - \frac{n^2-1}{12}\right]\lambda_2,$$

---

4 w sensie numerycznym

5 Istnieje również znacznie bardziej złożona metoda ogólna wielomianów ortogonalnych, w której nie żąda się spełnienia warunku (13.5), nie będziemy jej jednak tu omawiać.

$$\xi_3(x) = \left[k^3(x) - \frac{3n^2-7}{20} k(x)\right] \lambda_3 , \tag{13.6}$$

$$\xi_4(x) = \left[k^4(x) - \frac{3n^2-13}{14} k^2(x) - \frac{3(n^2-1)(n^2-9)}{560}\right] \lambda_4 .$$

gdzie

$$k(x) = \frac{1}{h}(x_i - \bar{x}) .$$

n jest liczbą wartości x_i, a średnią $\bar{x}$ można uzyskać ze wzoru

$$\bar{x} = \frac{1}{2}(x_1 + x_n)$$

Współczynniki λ_1, λ_2, ... są dobierane w zależności od n tak, aby wielomiany $\xi_j(x)$ przyjmowały tylko wartości całkowite.

Wartości wielomianów ortogonalnych w punktach x_i spełniają warunki ortogonalności

$$\sum_{i=1}^{n} \xi_j(x_i)\, \xi_{j'}(x_i) = 0 \qquad \text{dla } j \neq j' \tag{13.7}$$

a ponadto

$$\sum_{i=1}^{n} \xi_j(x_i) = 0 \qquad \text{dla każdego } j.$$

Model regresji wielomianowej (13.3) możemy napisać teraz w równoważnej mu postaci modelu regresji wielokrotnej względem wielomianów $\xi_j(x)$:

$$m(x) = a_0 + a_1 \xi_1(x) + a_2 \xi_2(x) + \ldots + a_p \xi_p(x) \tag{13.8}$$

Mając funkcję regresji w postaci (13.8) po podstawieniu wzorów (13.6) i redukcji otrzymamy z powrotem postać (13.3) dogodną do posługiwania się w praktyce. Tak więc parametry a_0, a_1, ..., a_p jednoznacznie określają współczynniki b_0, b_1, ..., b_p.

W celu wyznaczenia ocen parametrów funkcji regresji (13.8) dane empiryczne $\{(x_i, y_i)\}$ przekształcamy na dane postaci $\{(\xi_{1i}, \xi_{2i}, ..., y_i)\}$. Wartości wielomianów ortogonalnych

ξ_{1i} , ξ_{2i} , ... obliczamy ze wzorów (13.6) lub też odczytujemy wprost z tablic statystycznych[6] wartości wielomianów i współczynników λ_j. Ze względu na warunki (13.7) otrzymujemy

$$\operatorname{cov}(\xi_j, \xi_{j'}) = 0 , \qquad \operatorname{cov}(\xi_j, y) = \sum_{i=1}^{n} \xi_{ji}\, y_i$$

oraz

$$\operatorname{var} \xi_j = \sum_{i=1}^{n} \xi_{ij}^2$$

stałe dla ustalonego n. Macierz kowariancji $\boldsymbol{C}$ ma postać

$$\boldsymbol{C} = \begin{bmatrix} \operatorname{var}\xi_1 & 0 & \dots & 0 \\ 0 & \operatorname{var}\xi_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \operatorname{var}\xi_p \end{bmatrix}$$

Zatem rozwiązania układu równań normalnych $\boldsymbol{C}\,\hat{\boldsymbol{a}} = c_y$ są następujące:

$$\hat{a}_j = \frac{\operatorname{cov}(\xi_j, y)}{\operatorname{var}\xi_j} \qquad (j = 1, 2, ..., p) \qquad \text{oraz } \hat{a}_0 = \bar{y} .$$

Ze względu na diagonalną postać macierzy $\boldsymbol{C}$ analizę wariancji w regresji można w tym przypadku uszczegółowić przypisując każdemu składnikowi funkcji regresji (każdemu wielomianowi ortogonalnemu) niezależną sumę kwadratów odchyleń:

$$\operatorname{var} R_j = \hat{a}_j \operatorname{cov}(\xi_j, y) = \frac{\operatorname{cov}^2(\xi_j, y)}{\operatorname{var}\xi_j}$$

oczywiście globalna zmienność regresji wynosi:

$$\operatorname{var} R = \sum_{j=1}^{p} \operatorname{var} R_j$$

---

6 załączonych na końcu skryptu

Odpowiedni schemat analizy wariancji w regresji wielomianowej według wielomianów ortogonalnych zawiera poniższa tabela (symbolem var E oznaczono zmienność resztową błędu).

Błędy współczynników regresji $\hat{a}_j$ obliczamy teraz ze wzoru

Źródło zmienności	Liczba stopni swobody	Suma kwadratów	Średni kwadrat	F_{emp}
Regresja w tym	p	var R	s_R^2	s_R^2/s_e^2
regr. liniowa	1	var R_1	$s_{R_1}^2$	$s_{R_1}^2/s_e^2$
regr. kwadrat.	1	var R_2	$s_{R_2}^2$	$s_{R_2}^2/s_e^2$
.....	...	.....	...	.....
Odchylenia od regresji	$n-p-1$	var E	s_e^2	—
Ogółem	$n-1$	var y	—	—

$$s_{\hat{a}_j} = \sqrt{\frac{s_e^2}{\text{var}\,\xi_j}} \;;$$

nie ma jednak potrzeby sprawdzania żadnej z hipotez szczegółowych

$$H_{0j}: a_j = 0$$

gdyż są one sprawdzane w analizie wariancji. Hipotezę H_{0j} odrzucamy, gdy

$$F_{\text{emp}}^{(j)} = \frac{s_{R_j}^2}{s_e^2} > F_{\alpha,1,n-p-1}$$

Gdy rozpatrujemy regresję wielomianową, zmiennymi niezależnymi będą kolejne potęgi x, istnieje więc naturalne uszeregowanie tych zmiennych. Wprowadzając zatem kolejne potęgi zmiennej x, można zbadać na podstawie schematu analizy regresji podanego w tabeli, czy wzrost dokładności równania regresji jest statystycznie istotny. Należy jednak pamiętać, że jeżeli na przykład, dołączenie kwadratu zmiennej nie zwiększy statystycznie dokładności regresji, nie będzie to oznaczało, że wpływ następnej potęgi także nie będzie istotny.

Dlatego należałoby raczej określić równanie regresji wyższego stopnia, a następnie, po zbadaniu istotności współczynnika regresji przy najwyższej potędze, ewentualnie obniżyć stopień równania.

Przykład 1. Predykcja pojemności czaszki[7]

Zadanie polega na oszacowaniu pojemności czaszki w sytuacji, gdy czaszka jest połamana lub uszkodzona i pojemność nie może być zmierzona bezpośrednio. W takim przypadku pojemność czaszki można z pewną dokładnością oszacować na podstawie wyników pomiarów pewnych cech zewnętrznych, jeżeli tylko takie pomiary są możliwe do wykonania. Podstawowe dane, które są tu potrzebne, to kompletne wyniki pomiarów zarówno pojemności czaszek jak i danych cech zewnętrznych wykonanych na kilku dobrze zachowanych czaszkach. Dane te posłużą nam do oszacowania funkcji regresji, która z kolei będzie służyła do predykcji.

Trzy podstawowe cechy zewnętrzne, na podstawie których można szacować pojemność C czaszki, to wymiar strzałkowy (odległość gładzizna — potylica) L, największa szerokość kości ciemieniowej B oraz wysokość (odległość pomiędzy podstawą czaszki i bregmą) H. Ponieważ wielkością, którą mamy szacować, jest pojemność, odpowiednią funkcją regresji może być funkcja

$$C = \gamma L^{\beta 1} B^{\beta 2} H^{\beta 3} \tag{13.9}$$

gdzie γ, β_1, β_2, β_3 są parametrami, które należy wyznaczyć. Wprowadzając nowe zmienne

$$y = \log_{10} C\,, \qquad x_1 = \log_{10} L\,, \qquad x_2 = \log_{10} B\,, \qquad x_3 = \log_{10} H\,,$$

możemy powyższą funkcję napisać w postaci

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3\,, \tag{13.10}$$

gdzie $\beta_0 = \log_{10}\gamma$. Warto zauważyć, że wybór postaci funkcji regresji jest w tym przypadku arbitralny, równie dobrze można przyjąć funkcję regresji postaci

$$C = \beta_0 + \beta_1 L + \beta_2 B + \beta_3 H \tag{13.11}$$

lub też

---

7 Zaczerpnięty z: Rao C. R.: Modele liniowe statystyki matematycznej, PWN, Warszawa, 1982

$$C = \gamma (LBH)^{\beta} \tag{13.12}$$

Jeśli nie mamy żadnych przesłanek interpretacyjnych (w tym przypadku tak nie jest!), to musimy zbadać jak zachowują się wszystkie trzy funkcje tzn. (13.9), (13.11), (13.12) w procesie predykcji i wybrać tę z nich, która jest w jakimś sensie dokładniejsza. Niewłaściwy wybór funkcji może prowadzić do bardzo niedokładnej predykcji, co zmusiłoby nas do wypróbowania szeregu alternatywnych rozwiązań na danym materiale doświadczalnym zanim znaleźlibyśmy dostatecznie umotywowany wzór.

Jeżeli rozważana funkcja regresji zależy liniowo od nieznanych stałych, jak to ma miejsce we wzorze (13.10), dla oszacowania tych stałych można zastosować metodę najmniejszych kwadratów. Przepiszmy wzór (13.10) w postaci

$$y = \beta_0' + \beta_1 (x_1 - \bar{x}_1) + \beta_2 (x_2 - \bar{x}_2) + \beta_3 (x_3 - \bar{x}_3)$$

gdzie $\bar{x}_1$, $\bar{x}_2$, $\bar{x}_3$ są średnimi wyników pomiarów x_1, x_2, x_3 natomiast

$$\beta_0' = \beta_0 - \beta_1 \bar{x}_1 - \beta_2 \bar{x}_2 - \beta_3 \bar{x}_3 \ .$$

Przypuśćmy, że dysponujemy wynikami (y_r, x_{1r}, x_{2r}, x_{3r}) pomiarów n czaszek. Jak wiadomo z rozdziału dotyczącego regresji wielokrotnej układ równań związany z minimalizacją wielkości

$$\sum_{r=1}^{n} [y_r - \beta_0' - \beta_1 (x_{1r} - \bar{x}_1) - \beta_2 (x_{2r} - \bar{x}_2) - \beta_3 (x_{3r} - \bar{x}_3)]^2$$

ma postać

$$\boldsymbol{S}\,\boldsymbol{b} = s_y$$

gdzie

$$\boldsymbol{S} = [S_{ij}] , \qquad S_{ij} = \sum_r (x_{ir} - \bar{x}_i)(x_{jr} - \bar{x}_j) \ ,$$

$$s_y = \sum_r (y_r - \bar{y})(x_{ir} - \bar{x}_i) \ ,$$

natomiast $\boldsymbol{b} = [b_1, b_2, b_3]^T$ jest wektorem oszacowań parametrów β_1, β_2, β_3.

W celu oszacowania funkcji regresji wykorzystano wyniki pomiarów $n = 86$ czaszek z Farringdon Street i otrzymano następujące wielkości

$$\bar{y} = 3{,}1685 \qquad \bar{x}_1 = 2{,}2752, \qquad \bar{x}_2 = 2{,}1523, \qquad \bar{x}_3 = 2{,}1128,$$

$$S = \begin{bmatrix} 0{,}01875 & 0{,}00848 & 0{,}00684 \\ 0{,}00848 & 0{,}02904 & 0{,}00878 \\ 0{,}00684 & 0{,}00878 & 0{,}02886 \end{bmatrix}, \qquad s_y = \begin{bmatrix} 0{,}03030 \\ 0{,}04410 \\ 0{,}03629 \end{bmatrix},$$

oraz $S_{yy} = 0{,}12692$. Są to wstępne rachunki niezbędne do dalszej analizy.

Macierz odwrotna do S jest równa

$$S^{-1} = \begin{bmatrix} 64{,}21 & -15{,}57 & -10{,}49 \\ -15{,}57 & 41{,}01 & -9{,}00 \\ -10{,}49 & -9{,}00 & 39{,}88 \end{bmatrix}$$

Na podstawie pierwszego wiersza tej macierzy otrzymujemy oszacowanie b_1

$$b_1 = 64{,}21 S_{1y} - 15{,}57 S_{2y} - 10{,}49 S_{3y} = 0{,}878$$

Analogicznie otrzymujemy

$$b_2 = 1{,}041, \qquad b_3 = 0{,}733,$$

i w końcu

$$b_0' = \bar{y} = 3{,}1685, \qquad \text{a stąd} \quad b_0 = -2{,}618 \ .$$

Powracając do wyjściowych zmiennych otrzymujemy następujące równanie regresji

$$C = 0{,}00241\, L^{0{,}878}\, B^{1{,}041}\, H^{0{,}733} \tag{13.13}$$

UWAGA: Pojemność czaszek była mierzona przez szczelne wypełnienie ich nasionami gorczycy, a następnie zważenie tych nasion. Wzór (13.13) może być stosowany w sposób ścisły tylko do predykcji pojemności zdefiniowanej w ten właśnie sposób.

Resztowa suma kwadratów jest równa

$$S_e = S_{yy} - \sum b_i S_{iy} = 0{,}12692 - 0{,}09911 = 0{,}02781.$$

Oszacowaniem współczynnika korelacji wielokrotnej, mierzącego stopień dokładności predykcji, czyli stopień zależności pomiędzy zmienną zależną i zmiennymi niezależnymi, jest

$$R^2 = \frac{S_{yy} - S_e}{S_{yy}} = \frac{0{,}09911}{0{,}12692} = 0{,}7809 \ ,$$

zatem

$$R = 0{,}8837 \ .$$

Oszacowaniem wariancji resztowej jest

$$s_e^2 = \frac{S_e}{n-p-1} = \frac{0{,}02781}{82} = 0{,}0003391 \ ,$$

Wariancje i kowariancje estymatorów b_i wyrażają się wzorami

$$\text{var } b_i = S^{ii} s_e^2 \qquad \text{oraz} \qquad \text{cov } (b_i, b_j) = S^{ij} s_e^2 \ ,$$

gdzie S^{ii}, S^{ij} są odpowiednimi elementami macierzy S^{-1}. Natomiast błędy standardowe cząstkowych współczynników regresji wyrażają się wzorami

$$s_{b_i} = \sqrt{S^{ii} s_e^2}$$

Wyznaczenie tych wartości zakończy estymację nieznanych parametrów i ich wariancji metodą najmniejszych kwadratów.

Rozpatrzymy teraz kilka zagadnień związanych z wyliczoną właśnie funkcją regresji i z jej wykorzystaniem.

1. Czy zmienne zależna i niezależna są istotnie ze sobą powiązane? W języku wprowadzonych wyżej oznaczeń hipoteza brzmi

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0 \ .$$

Hipoteza ta jest równoważna hipotezie, że prawdziwy współczynnik korelacji wielokrotnej jest równy zeru.

Posłużymy się techniką analizy wariancji. Resztową sumę kwadratów S_e obliczyliśmy już poprzednio i otrzymaliśmy wynik 0,02781; liczba stopni swobody wynosi $(n - p - 1) = 82$. Jeżeli hipoteza H_0 jest prawdziwa, to regresji odpowiada suma kwadratów

$$S_{reg} = S_{yy} - S_e = 0{,}12692 - 0{,}02781 = 0{,}09911 \ ,$$

a odpowiadająca jej liczba stopni swobody jest równa $\nu = 3$. Analizę sum kwadratów przedstawiono w poniższej tabeli.

Zauważmy, że iloraz wariancji z tablicy można łatwo obliczyć na podstawie wartości R^2 (kwadratu współczynnika korelacji wielokrotnej), n (liczebności próby) oraz p (liczby zmiennych niezależnych) za pomocą wzoru

$$F = \frac{R^2}{1-R^2} \cdot \frac{n-p-1}{P} = \frac{0{,}7809}{1-0{,}7809} \cdot \frac{86-3-1}{3} = 97{,}41$$

	Stopnie swobody	Suma kwadratów	Średni kwadrat	F
Regresja	3	0,09911	0,033037	97,41
Reszta	82	0,02781	0,0003391	
Ogółem	85	0,12692		

Wynika stąd, że rozważaną hipotezę można testować tylko na podstawie zaobserwowanej wartości współczynnika korelacji wielokrotnej. Iloraz wariancji jest równy 74.41 przy $\nu_1 = 3$ i $\nu_2 = 82$ stopniach swobody i jest istotny na poziomie 1%, co wskazuje na przydatność rozważanych zmiennych niezależnych dla predykcji.

2. A oto inne hipotezy związane z rozpatrywanym modelem. Specjaliści w zakresie ewolucji przypuszczają, że szerokość czaszki rośnie stosunkowo szybciej niż inne jej parametry. Można zatem zbadać, czy wszystkie trzy zmienne niezależne mają takie samo znaczenie dla predykcji, gdyż na podstawie wykonanych oszacowań można zauważyć, że współczynnik b_2 przy największej szerokości kości ciemieniowej B jest wyższy od każdego z pozostałych współczynników regresji. Odpowiednia hipoteza będzie miała postać

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta \ .$$

Innego rodzaju hipoteza związana jest z faktem, że dla predykcji pojemności czaszki używa się czasami prostego wzoru postaci

$$C = \beta_0' \ LBH \ .$$

Weryfikacja adekwatności takiego wzoru jest równoważna weryfikacji hipotezy

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 1 \ .$$

Przekonawszy się z kolei, że poszczególne współczynniki β różnią się od jedności możemy zbadać interesujące zagadnienie, czy wszystkie one sumują się do 3 i tylko nierównomiernie rozkładają się pomiędzy trzy badane wymiary czaszki. Polegałoby to na weryfikacji hipotezy

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 3 \ .$$

Proponujemy Czytelnikowi zastanowienie się nad sposobem testowania opisanych hipotez.

3. Często pożądane jest sprawdzić, czy wprowadzenie dodatkowej zmiennej zwiększy dokładność predykcji. Na przykład, w rozważanym wyżej zadaniu możemy testować, czy konieczne jest dołączenie zmiennej H do zmiennych L i B. Jest to równoważne weryfikacji hipotezy, że $\beta_3 = 0$. Estymator tego współczynnika regresji jest równy $b_3 = 0{,}733$, a jego wariancja jest równa $\text{var}\, b_3 = S^{33} s_e^2$. Odpowiedni iloraz, przy $\nu_1 = 1$ i $\nu_2 = 82$ stopniach swobody, wynosi

$$F = \frac{b_3^2}{s^{33} s_e^2} = \frac{(0{,}733)^2}{39{,}88} \cdot \frac{1}{0{,}0003391} = \frac{0{,}01347}{0{,}0003391} = 39{,}72$$

Wartość tego ilorazu jest istotna na poziomie 1%, co wskazuje na to, że zmienna H odgrywa również istotną rolę. Gdyby b_3 nie było istotne, sumę kwadratów związaną z tym współczynnikiem, a mianowicie

$$\frac{b_3^2}{S^{33}} = 0{,}01347$$

można by dodać do resztowej sumy kwadratów $S_e = 0{,}02781$; otrzymalibyśmy sumę kwadratów 0,04128 opartą na 83 stopniach swobody i wtedy oszacowaniem s_e^2 byłoby 0,0004973. Po położeniu $b_3 = 0$ należałoby ponownie obliczyć najlepsze estymatory b_1 i b_2 współczynników regresji, opisanej tym razem równaniem

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

4. Skonstruowany wyżej wzór dla predykcji może być również przydatny wtedy, gdy chcielibyśmy oszacować średnią pojemność czaszki na podstawie próby, w której udało się zmierzyć tylko wielkości L, B oraz H. Estymację taką można wykonać na dwa sposoby: można oszacować pojemność każdej czaszki w próbie i obliczyć średnią z uzyskanych wyników, albo zastosować wzór na predykcję do średniej liczb L, B i H zaobserwowanych w próbie. Interesujące jest, czy obie metody prowadzą do tego samego wyniku. W celu uzyskania odpowiedzi na to pytanie oszacowano średnią pojemność czaszki w dodatkowej próbie 29 czaszek z Farringdon Street; dla każdej z tych czaszek zmierzono L, B oraz H, ale pomiar C nie był możliwy.

   Dla L, B oraz H otrzymano, odpowiednio, średnie 191,1, 143,1 i 129,0. Na podstawie wzoru

$$C = 0{,}00241\, L^{0{,}878}\, B^{1{,}041}\, H^{0{,}739}$$

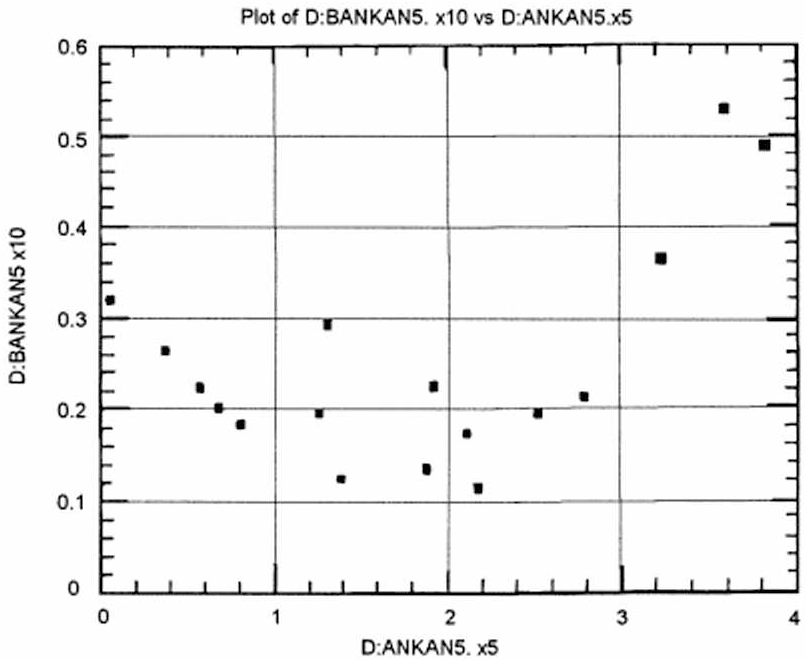
otrzymano dla C wartość 1498,3. Szacując C dla każdej z czaszek oddzielnie i obliczając średnią, otrzymano 1498,2.

Taką samą estymację przeprowadzono na podstawie 22 męskich czaszek z Moorfields. W tym przypadku dostępne były wyniki pomiarów wszystkich czterech parametrów. Dla L, B oraz H otrzymano średnie 189,5, 142,5 oraz 128,8, co dało wartość 1479,0 dla średniej C. Szacując C indywidualnie dla każdej czaszki i obliczając średnią otrzymano 1480,0. Powyższe wyniki sugerują, że obie metody dają bardzo podobne oszacowania.

Przykład 2. Regresja kwadratowa

Analizując rys. 13.1 przedstawiający zależność zmiennej $x10$ od $x5$[8] widzimy, że zależność ta jest wyraźnie nieliniowa, zbliżona do kwadratowej. Postaramy się zatem dopasować do danych pomiarowych krzywą postaci:

$$x10 = \beta_0 + \beta_1 x5 + \beta_2 x5^2$$



Rys. 13.1 Zależność $x10 = f(x5)$

8 tzn. jod związany w białku proteinowym przed rozpoczęciem i po zakończeniu terapii u chorych z zaburzeniem tarczycy należy do 1-szej klasy (liczebność klasy $n = 16$) — por. Dodatek 1

Zmienną $x10$ uważać będziemy za zmienną zależną, natomiast $x5$ za zmienną niezależną.

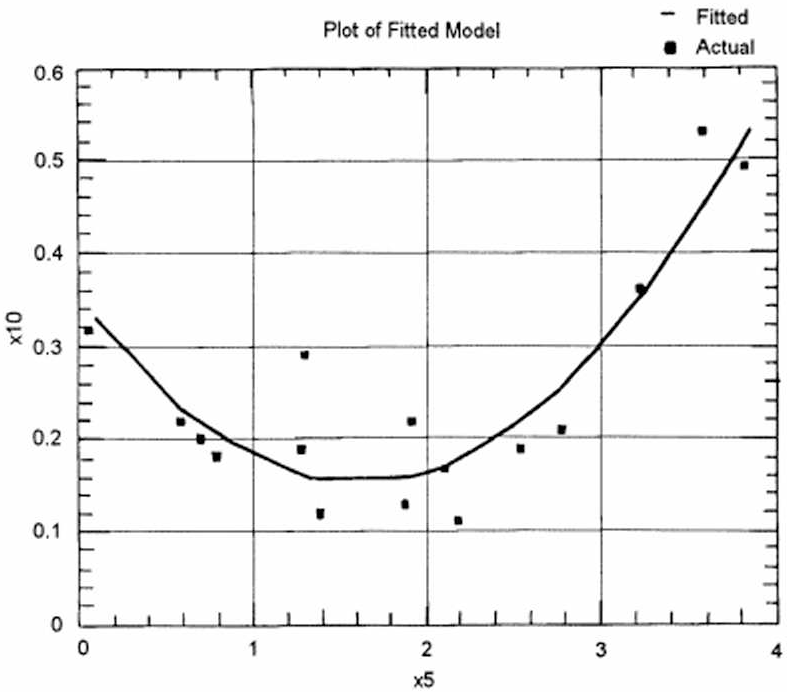
Po dokonaniu przeliczeń metodą najmniejszych kwadratów otrzymujemy następujące oszacowania współczynników równania:

Współczynnik	Estymator	Błąd standard.
β_0	0,3502	0,04416
β_1	–0,2440	0,04925
β_2	0,0755	0,01185

czyli

$$x10 = 0{,}3502 - 0{,}2440x5 + 0{,}0755x5^2$$

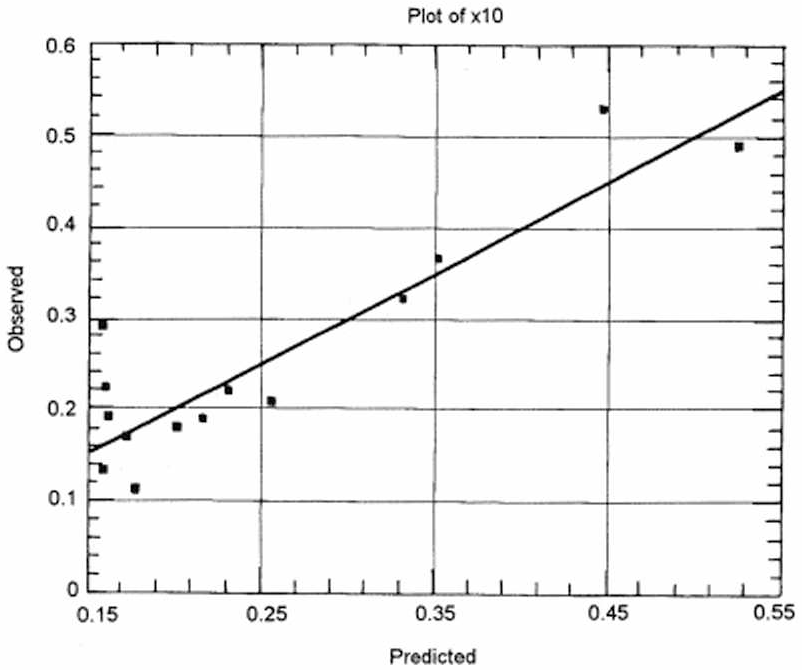
przy czym wszystkie współczynniki są istotne. Rysunek 13.2 pokazuje wykres dobranej do danych empirycznych krzywej, natomiast na rysunku 13.3 pokazano zależność między danymi zaobserwowanymi zmiennej $x10$ a wartościami wyliczonymi (tzn. predykcją).



Rys. 13.2 Wykres krzywej empirycznej dopasowanej do danych pomiarowych

Poniżej podajemy jeszcze globalną tabelę analizy wariancji dla przyjętego modelu kwadratowego — jak widać zależność jest bardzo istotna.

Źródła zmienności	Liczba stopni swobody	Suma kwadratów	Średni kwadrat	F
Regresja	2	1,15458	0,57729	186,2
Odchylenie od regresji	13	0,04032	0,00310	
Ogółem	15	1,19490		



Rys. 13.3 Wykres zależności między zaobserwowanymi wartościami zmiennej x_{10} a wartościami wyliczonymi.

14. ANALIZA KANONICZNA

Współczynnik korelacji prostoliniowej rozpatrywany w pierwszej części skryptu mierzy zależność między dwiema zmiennymi losowymi. Omawiany w tej części skryptu współczynnik korelacji wielokrotnej mierzy z kolei zależność między jedną zmienną losową i układem innych zmiennych. W rzeczywistości jest to maksymalny współczynnik korelacji między daną zmienną a kombinacją liniową pozostałych zmiennych — a zatem jest to współczynnik korelacji między daną zmienną i jej najlepszą (tu: w sensie minimum błędu średniokwadratowego) predykcją zbudowaną na tych pozostałych zmiennych.

Analiza kanoniczna zajmuje się uogólnieniem koncepcji korelacji w przypadku współzależności pomiędzy dwoma zbiorami zmiennych losowych. Innymi słowy, w ramach analizy kanonicznej szuka się odpowiedzi na pytanie jaki jest zakres równoczesnego wpływu całego zbioru zmiennych niezależnych $x = x_1, x_2, ..., x_p$ na cały zbiór zmiennych zależnych $y = y_1, y_2, ..., y_q$?

Odpowiedzią jest pewien syntetyczny wskaźnik statystyczny, informujący o zakresie determinacji (tzn. określania) zbioru zmiennych x względem zbioru zmiennych y, jest to liczbowa miara „korelacji" między dwoma zbiorami zmiennych o różnej ich liczbie.

Korzystając z metod analizy kanonicznej można również uzyskać odpowiedź na bardziej szczegółowe pytania, takie jak:

— który zbiór zmiennych niezależnych, spośród kilku możliwych wyjaśnia największy zakres zmienności (wariancji) w obrębie zbioru y?

— czy wprowadzenie nowych zmiennych niezależnych lub zależnych, przy zachowaniu wyjściowej struktury zbiorów x i y, zwiększa zakres wyjaśnionej wariancji całkowitej?

— które zmienne przestrzeni x, rozpatrywane łącznie, wyjaśniają największy zakres wariancji dla zmiennych przestrzeni y, również rozpatrywanych łącznie?

Rozważmy układ dwóch zbiorów zmiennych losowych $[x, y]$, gdzie
$x = [x_1, x_2, ..., x_p]^T$ jest wektorem zmiennych objaśniających,
$y = [y_1, y_2, ..., y_q]^T$ — wektorem zmiennych objaśnianych.

Interesuje nas złożony związek między zbiorami x i y, a mówiąc ściślej: chcemy za pomocą zmiennych x prognozować zmienne y.

Macierz kowariancji układu $[x, y]$ można wyrazić w postaci blokowej jako

$$\Sigma = \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix}, \qquad \text{gdzie } \Sigma_{yx}^{T} = \Sigma_{xy} \tag{14.1}$$

gdzie macierz Σ jest macierzą symetryczną stopnia $p + q$. Zakładamy również, że Σ jest macierzą nieosobliwą, przy czym jej wyznacznik $|\Sigma| > 0$. Oznaczmy przez s rząd podmacierzy Σ_{xy}.

Szukamy pary funkcji liniowych najlepiej dopasowanej (reprezentującej) obydwa zbiory zmiennych. Główna idea analizy kanonicznej polega na tym, aby funkcję liniową zmiennych x postaci

$$u = l_1 x_1 + l_2 x_2 + \ldots + l_p x_p$$

i funkcję liniową zmiennych y postaci

$$v = m_1 y_1 + m_2 y_2 + \ldots + m_q y_q$$

dobrać w taki sposób, aby korelacja w parze zmiennych kanonicznych była **maksymalna**. Spełnienie tego istotnego warunku oznacza, że pary zmiennych kanonicznych u_t i v_t możemy uważać za dobrą reprezentację danych wyjściowych w ramach przyjętego modelu. Liczba par zmiennych kanonicznych jest równa liczbie zmiennych zależnych, tzn. zmiennych objaśnianych.

Szukamy zatem funkcji liniowych $u = L^T x$ i $v = M^T y$ spełniających warunki:

1. $\sigma^2(u) = \sigma^2(v) = 1$,
2. kowariancja $\text{cov}(u, v) = \max$.

Funkcje u i v noszą nazwę zmiennych kanonicznych, a wektory L i M — wektorów kanonicznych. Pierwszy warunek z wyżej wymienionych oznacza unormowanie wariancji szukanych funkcji liniowych, przy czym

$$\sigma^2(u) = \sigma^2(L^T x) = (L^T x)(L^T x)^T = L^T x x^T L = L^T \Sigma_{xx} L ,$$

i analogicznie

$$\sigma^2(v) = \sigma^2(M^T x) = (M^T y)(M^T y)^T = M^T y y^T M = M^T \Sigma_{yy} M .$$

Natomiast kowariancja

$$\text{cov}(u, v) = \rho_{uv}$$

jest (również z uwagi na unormowane wariancje) równa współczynnikowi korelacji zmiennych u i v, przy czym zachodzi równość

$$\rho_{uv} = L^T \Sigma_{xy} M \tag{14.2}$$

Maksymalnej korelacji (warunek 14.2. przy warunku ubocznym 14.1.) poszukuje się metodą nieoznaczonych mnożników Lagrange'a. Po przekształceniach, (które tu pomijamy) otrzymuje się układ równań jednorodnych

$$\left[\Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx} - \rho^2 \Sigma_{xx}\right] L = 0 \tag{14.3}$$

lub symetrycznie

$$\left[\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy} - \rho^2 \Sigma_{yy}\right] M = 0 \tag{14.4}$$

gdzie $\rho = \rho_{uv}$. Układ ten (tzn. (14.3)) ma niezerowe rozwiązanie tylko wtedy, gdy macierz układu jest osobliwa. Otrzymujemy zatem równanie wyznacznikowe

$$\left|\Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx} - \rho^2 \Sigma_{xx}\right| = 0 \tag{14.5}$$

które przy założeniu, że macierze Σ_{xx} i Σ_{yy} są nieosobliwe, sprowadza się do dobrze znanego równania charakterystycznego

$$|A - \rho^2 I| = 0$$

gdzie macierz $A = \Sigma_{xx}^{-1}\Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}$, lub symetrycznie według drugiego równania(tzn. 14.(4)) $A = \Sigma_{yy}^{-1}\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}$.

Pierwsze z równań ma p pierwiastków, natomiast drugie q pierwiastków. Ponieważ jednak niezerowe dodatnie pierwiastki obu równań są identyczne (i jest ich $s = rz\ \Sigma_{xy} \leq \min(p, q)$), zatem można je reprezentować tymi samymi symbolami. Oznaczmy je wobec tego przez $\rho_1, \rho_2, \ldots, \rho_s$ i uporządkujmy, tak aby zachodziła nierówność:

$$\rho_1 \geq \rho_2 \geq \ldots \geq \rho_s$$

Wielkości $\rho_1, \rho_2, ..., \rho_s$ noszą nazwę korelacji kanonicznych. Każdemu z tych pierwiastków odpowiada wektor kanoniczny[1] $L_t (t = 1, 2, ..., s)$. Zerowym pierwiastkom równania (14.5) odpowiadają wektory własne oznaczone L_t o numerach $s < t \leq p$. Mając wektor L_t, sprzężony z nim wektor M_t wyliczymy ze wzoru

$$M_t = \rho^{-1} \Sigma_{yy}^{-1} \Sigma_{yx} L_t \tag{14.6}$$

Zachodzi również symetryczna zależność

$$L_t = \rho^{-1} \Sigma_{11}^{-1} \Sigma_{12} M_t \tag{14.6a}$$

Wektory kanoniczne L_t i M_t określają zmienne kanoniczne u_t i v_t, przy czym

$$\operatorname{cov}(u_t, v_t) = \begin{cases} \rho_t & \text{dla } t = t' \\ 0 & \text{dla } t \neq t' \end{cases}$$

Jest to bardzo istotne, gdyż oznacza istnienie korelacji tylko między sprzężonymi zmiennymi kanonicznymi — wszystkie pozostałe korelacje są zerowe! Dodatkowo zmienne u_t oraz v_t są wewnętrznie nieskorelowanie, tzn.

$$\operatorname{cov}(u_t, u_{t'}) = 0 \qquad \text{dla } t \neq t'$$

i analogicznie

$$\operatorname{cov}(v_t, v_{t'}) = 0 \qquad \text{dla } t \neq t'$$

Innymi słowy, pierwsza zmienna kanoniczna w p-wymiarowej przestrzeni x jest skorelowana jedynie z pierwszą zmienną kanoniczną w q-wymiarowej przestrzeni y, druga zmienna kanoniczna w przestrzeni x jedynie z drugą zmienną kanoniczną w przestrzeni y, ... itd. Odpowiednie współczynniki korelacji są równe $\rho_1, \rho_2, ..., \rho_s$.
Jak wynika z tych ustaleń macierz kowariancji zmiennych kanonicznych u oraz v ma postać blokową:

---

1 czyli wektor własny macierzy A

$$\begin{bmatrix} I_s & R \\ R & I_s \end{bmatrix}$$

Dowodzi się, że zmienne kanoniczne są niezmiennicze ze względu na liniowe przekształcenia zmiennych pierwotnych x_i lub y_j. Stąd wniosek, iż jest obojętne, czy do obliczeń użyte zostaną zmienne pierwotne, czy też ich standaryzowane postaci o $E(x_i) = E(y_j) = 0$ i $\sigma^2(x_i) = \sigma^2(y_j) = 1$.

Analizując korelacje między zmiennymi pierwotnymi a kanonicznymi uzyskujemy związki[2]:

$$\operatorname{cov}(x_i, v_j) = \rho_j \operatorname{cov}(x_i, u_j) \qquad (i = 1, 2, ..., p;\ j = 1, 2, ..., s) \tag{14.7}$$

$$\operatorname{cov}(y_i, u_j) = \rho_j \operatorname{cov}(y_i, v_j) \qquad (i = 1, 2, ..., q;\ j = 1, 2, ..., s) \tag{14.8}$$

oraz

$$\left.\begin{aligned} \operatorname{cov}(x_i, u_i) &= \sum_{k=1}^{p} l_{ki} \operatorname{cov}(x_k, x_i) \qquad (\text{gdzie } L = [l_k]), \\ \operatorname{cov}(y_j, v_j) &= \sum_{k=1}^{q} m_{kj} \operatorname{cov}(y_k, y_j) \qquad (\text{gdzie } M = [m_k]). \end{aligned}\right\} \tag{14.9}$$

A zatem współczynniki korelacji zmiennych pierwotnych ze zmiennymi kanonicznymi tej samej grupy zmiennych (tzw. wewnętrzne korelacje kanoniczne) uzyskuje się z mnożenia wektora L dla x (lub M dla y) przez macierz korelacji zmiennych pierwotnych. Natomiast korelacje zmiennych pierwotnych jednej grupy ze zmiennymi kanonicznymi drugiej są równe iloczynom korelacji kanonicznej przez korelację wewnętrzną.

Suma kwadratów korelacji wewnętrznych równa się jedności, stąd:

$$\sigma^2(x_i) = \sum_{k=1}^{p} \operatorname{cov}^2(x_k, x_j) \qquad \text{i} \qquad \sigma^2(y_j) = \sum_{k=1}^{q} \operatorname{cov}^2(y_k, y_j)\ . \tag{14.10}$$

Para wzorów (14.10) pozwala w prosty sposób interpretować zmienne kanoniczne jako liniowe funkcje zmiennych pierwotnych determinujące (określające) w różnym stopniu

---

2 przy założeniu $\sigma^2(x_i) = \sigma^2(y_j) = 1$.

wariancje tych zmiennych wyjściowych. Miarą owej determinacji jest kwadrat współczynnika korelacji wewnętrznej.

Oszacujemy teraz parametry przyjętego modelu na podstawie próby danych empirycznych.

Niech

$$X = [x_{ik}] \qquad (i = 1, 2, ..., p;\ k = 1, 2, ..., n)$$

będzie macierzą obserwacji zmiennych x, a

$$Y = [y_{jk}] \qquad (j = 1, 2, ..., q;\ k = 1, 2, ..., n)$$

macierzą obserwacji zmiennych y. Na podstawie próby wyliczamy estymatory

$$\left.\begin{aligned} \hat{\Sigma}_{xx} &= S_{xx} = \frac{1}{n-1} X_0 X_0^T, \\ \hat{\Sigma}_{yy} &= S_{yy} = \frac{1}{n-1} Y_0 Y_0^T, \\ \hat{\Sigma}_{xy} &= S_{xy} = S_{yx}^T = \frac{1}{n-1} X_0 Y_0^T, \end{aligned}\right\} \tag{14.11}$$

gdzie X_0 oraz Y_0 są macierzami odchyleń obserwacji od średnich, tj. powstają przez odjęcie w macierzach X i Y odpowiednich średnich od każdego elementu; macierz

$$\hat{\Sigma} = \begin{bmatrix} S_{xx} & S_{xy} \\ S_{yx} & S_{yy} \end{bmatrix}$$

jest macierzą kowariancji danych empirycznych. Dodatkowo wprowadzamy oszacowanie $\hat{\rho}_t = r_t$.

Wektory kanoniczne z próby u_t i v_t ($t = 1, 2, ..., s$) znajdujemy wyznaczając wektory z próby L_t i M_t ze wzorów (14.3) oraz (14.6) po podstawieniu w nich estymatorów odpowiednich macierzy kowariancji. Pozostałe wzory — wcześniej napisane — pozostaną prawdziwe po zastąpieniu w nich parametrów modelu ich estymatorami.

Otrzymane zmienne kanoniczne z próby mogą służyć prognozowaniu. Tak więc dla prognozy wartości zmiennych y_j bierzemy zmienne kanoniczne u_i ($i = 1, 2, ..., s$) odpowiadające niezerowym korelacjom kanonicznym. Po odpowiednich przekształceniach otrzymamy

$$Y_0 = S_{yy} \hat{M} \hat{M}^T S_{yx} \hat{L} \hat{L}^T X_0, \qquad \text{gdzie} \quad \hat{L} = [\hat{L}_t],\ \hat{M} = [\hat{M}_t]. \tag{14.12}$$

Zmienne odpowiadające zerowym korelacjom kanonicznym (zerowym pierwiastkom równania charakterystycznego) są tu pominięte, ponieważ nie wnoszą do równania (14.12) żadnej informacji. Można wykazać, że równanie (14.12) jest macierzową postacią zestawu q równań regresji wielokrotnej kolejnych zmiennych y_j względem zbioru zmiennych $\{x_i\}$. Zatem wielkości

$$R^2_{y_j \cdot x} = \frac{\sum_{k=1}^{q} r_k \operatorname{cov}(v_k, y_j)}{\operatorname{var} y_j} \tag{14.13}$$

są współczynnikami determinacji y_j przez zmienne $\{x_i\}$[3].
Natomiast średnia ważona współczynników determinacji danych wzorem (14.13), czyli

$$R^2_{y \cdot x} = \frac{\sum_{j=1}^{q} \operatorname{var} y_j * R^2_{y_j \cdot x}}{\sum_{j=1}^{q} \operatorname{var} y_j} \tag{14.14}$$

nosi nazwę złożonego współczynnika determinacji i jest miarą przeciętnej determinacji zespołu zmiennych y_j przez zespół zmiennych x_i. Pierwiastek ze złożonego współczynnika determinacji tzn. $R^2_{y \cdot x}$ jest nazywany przez niektórych autorów złożonym współczynnikiem korelacji.
Wzór (14.14) można również zapisać w postaci

$$R^2_{y \cdot x} = \frac{tr\,(S_{22}\hat{M}\,(\hat{L}^T S_{12}\hat{M})^2\,\hat{M}^T S_{22})}{tr\,(S_{22})},$$

gdzie tr(A) oznacza ślad macierzy A.
Złożony współczynnik determinacji bywa również nazywany współczynnikiem redundancji lub też złożoną determinacją.

Omówienia wymaga jeszcze sposób testowania hipotez w analizie korelacji kanonicznych. Załóżmy w tym celu, że wektory losowe x oraz y podlegają rozkładowi normalnemu. Hipotezą globalną interesującą nas w przypadku wyodrębniania dwóch zbiorów zmiennych jest hipoteza

---

3 Są to identyczne współczynniki determinacji jak w przypadku korelacji wielorakiej

H_0 : wektory x i y są niezależne.

Jest ona równoważna hipotezie:

$$H_0 : \sum \rho_i^2 = 0 \ ,$$

tzn. że wszystkie korelacje kanoniczne w przypadku normalności rozkładów są zerowe. Istnieje kilka testów służących sprawdzeniu tej hipotezy. Statystyką opartą o iloraz wiarygodności jest wielkość

$$\Lambda_{emp} = \prod_{t=1}^{s} (1 - r_t^2) \ ,$$

mająca rozkład Wilksa przy założeniu prawdziwości hipotezy H_0 z parametrami $n - 1$, p, q. W szczególnych przypadkach, gdy $p = 1$ lub 2 albo też $q = 1$ lub 2, rozkład Λ_{emp} wyraża się przez centralny rozkład F w poniższy sposób

$$\frac{n-1}{q} * \frac{1-\Lambda}{\Lambda} \sim F_{q,\,n-1} \quad \text{dla} \quad p = 1$$

$$\frac{n-2}{q} * \frac{1-\sqrt{\Lambda}}{\sqrt{\Lambda}} \sim F_{2q,\,2n-2} \quad \text{dla} \quad p = 2$$

$$\frac{n-p}{p} * \frac{1-\Lambda}{\Lambda} \sim F_{p,\,n-p} \quad \text{dla} \quad s = 1$$

$$\frac{n-p}{p} * \frac{1-\sqrt{\Lambda}}{\sqrt{\Lambda}} \sim F_{2p,\,2n-p} \quad \text{dla} \quad s = 2$$

Dla innych wartości p oraz q używa się zmiennej

$$\chi^2_{emp} = -a \log \Lambda_{emp} \ , \qquad \text{gdzie} \quad a = n - 1 + \frac{1}{2}(p - q + 1) \tag{14.15}$$

mającej dla dużych n rozkład zbliżony do χ^2, albo też zmiennej

$$F_{emp} = \frac{1-\sqrt{\Lambda}}{p\sqrt{\Lambda}} * \frac{ab - 2c}{pq} \ , \tag{14.16}$$

$$\text{gdzie } b = \frac{p^2 q^2 - 4}{p^2 + q^2 - 5} \ , \qquad c = \frac{pq - 2}{4} \ ,$$

której rozkład można aproksymować rozkładem F dla $\nu_1 = pq$ i $\nu_2 = ab - 2c$ stopni swobody.

Odrzucenie hipotezy globalnej H_0 prowadzi do hipotez o liczbie zmiennych kanonicznych. Hipotezy te testujemy sekwencyjnie. Hipoteza, że istotna jest tylko jedna zmienna kanoniczna, równoważna jest hipotezie

$$H_{01} : \rho_2 = \rho_3 = \ldots = 0 \ .$$

Po jej odrzuceniu testujemy hipotezę

$$H_{02} : \rho_3 = \rho_4 = \ldots = 0 \ .$$

Stawiamy taką serię hipotez dopóty, dopóki nie pojawi się hipoteza do przyjęcia na danym poziomie istotności α. Możemy się również posłużyć statystyką Wilksa

$$\Lambda_{emp}^{(k)} = \prod_{i=k+1}^{s} (1 - r_i^2)$$

o parametrach $n - k$, $p - k$, $q - k$. Przy obliczaniu wartości krytycznej testu posługujemy się też aproksymacjami (14.15) lub (14.16) podstawiając odpowiednio $n - k$ w miejsce n, $p - k$ w miejsce p oraz $q - k$ w miejsce q.

Przykład 1.

Posłużymy się danymi dotyczącymi osób z nadczynnością gruczołów tarczycowych (dodatek 2). Weźmiemy pod uwagę 16-to osobową grupę wyleczonych pacjentów. Załóżmy, że pierwszy zbiór zmiennych $x = [x3, x5]^T$ obejmuje dwie zmienne niezależne, natomiast drugi zbiór $y = [x8, x10]^T$ zawiera również dwie zmienne zależne.

Wprowadzamy dwie zmienne kanoniczne u_1 i u_2 reprezentujące zbiór x i dwie zmienne kanoniczne v_1 i v_2 reprezentujące zbiór y. Zmienne zależne i niezależne przekształcamy na zmienne kanoniczne u_i oraz v_i w dwóch wymiarach. Otrzymujemy zatem dwa równania dla u:

$$u_1 = 0{,}75926x3 + 0{,}84151x5$$
$$u_2 = 0{,}68972x3 - 0{,}58656x5$$

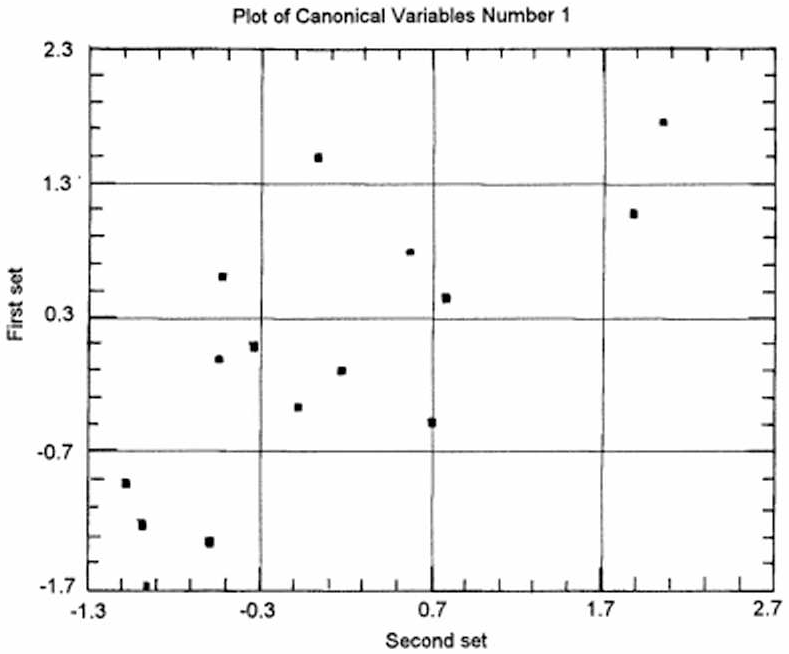
oraz dwa równania dla zmiennych kanonicznych v:

$$v_1 = 0{,}33784x8 + 0{,}79339x10$$
$$v_2 = 1{,}08742x8 - 0{,}81680x10$$

Zmienne kanoniczne u_i oraz v_i są funkcjami liniowymi x i y tak dobranymi, aby korelacje między u_i a v_i osiągnęły wartość maksymalną. W naszym przypadku otrzymujemy

$$\text{cov}(u_1, v_1) = r_{11} = 0{,}7026$$
$$\text{cov}(u_2, v_2) = r_{22} = 0{,}3362$$

Jeszcze raz podkreślamy, że współczynnika korelacji kanonicznej nie interpretuje się tak samo jak innych „normalnych" współczynników korelacji. Współczynnik korelacji kanonicznej informuje jedynie o tym, w jakim stopniu udało się maksymalnie skorelować odpowiednią parę zmiennych kanonicznych. Wartości współczynników korelacji kanonicznej podane powyżej świadczą o tym, że funkcje liniowe słabo reprezentują dane wejściowe, albo, brak współzależności między zbiorami x i y.
Na rys. 14.1 przedstawiono dodatkowo pozycję każdej osoby badanej, przy czym zmienne kanoniczne u_1 i v_1 potraktowano jako współrzędne, tzn. $u_1 = f(v_1)$.



Rys. 14.1 Rozkład obserwacji w przestrzeni $u_1 = f(v_1)$ dla niezmodyfikowanych zbiorów wejściowych

Jeżeli zmodyfikujemy teraz nasze zbiory danych tak, aby zawierały one po 4 zmienne:

$$x = [x1, x2, x3, x4]$$
$$y = [x6, x7, x8, x9] \ ,$$

o otrzymamy następujące równania dla czterech par zmiennych kanonicznych:

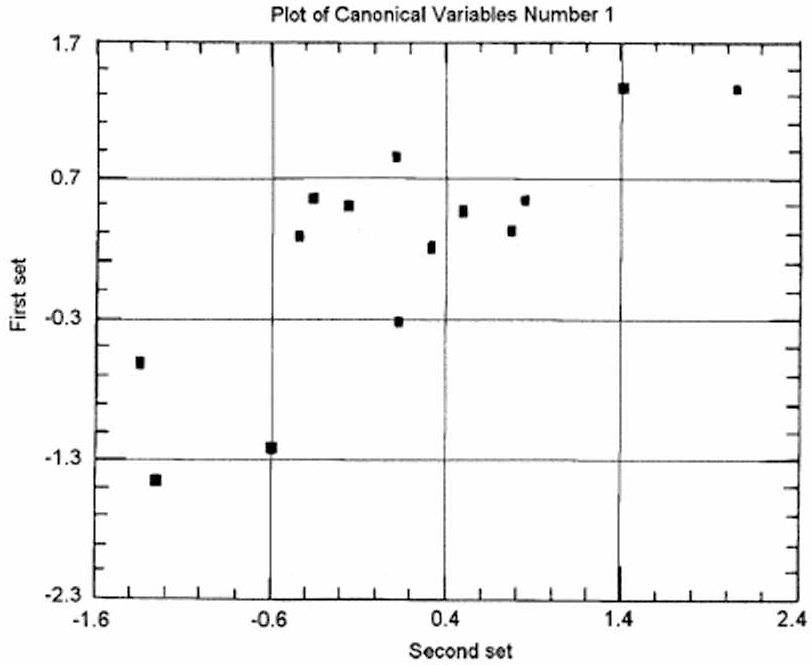
$$u_1 = 0{,}80x1 - 0{,}28x2 + 0{,}21x3 + 0{,}44x4$$
$$u_2 = -0{,}08x1 - 1{,}83x2 + 1{,}77x3 - 0{,}03x4$$
$$u_3 = -0{,}52x1 + 0{,}02x2 + 1{,}96x3 - 1{,}51x4$$
$$u_4 = 0{,}99x1 - 1{,}23x2 + 0{,}84x3 - 0{,}91x4$$

oraz dla v_i:

$$v_1 = -1{,}36x6 + 2{,}43x7 + 1{,}87x8 - 2{,}15x9$$
$$v_2 = -2{,}55x6 + 0{,}48x7 + 1{,}33x8 + 0{,}69x9$$
$$v_3 = -0{,}98x6 + 2{,}00x7 - 2{,}47x8 + 1{,}66x9$$
$$v_4 = 2{,}79x6 - 3{,}97x7 - 1{,}17x8 + 2{,}98x9$$

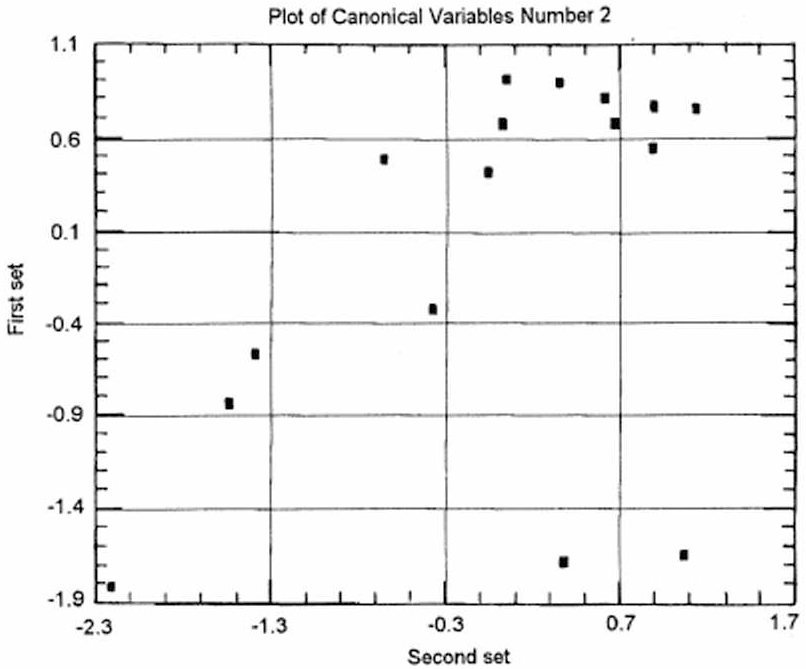
Korelacje kanoniczne wynoszą w tym przypadku

$$r_{11} = 0{,}7329, \quad r_{22} = 0{,}4585,$$
$$r_{33} = 0{,}2108, \quad r_{44} = 0{,}0484.$$



Rys. 14.2 Rozkład obserwacji w przestrzeni $u_1 = f(v_1)$.

Na rys. 14.2 przedstawiono, tak jak poprzednio, pozycję każdej osoby badanej w układzie współrzędnych $u_1 = f(v_1)$, natomiast na rys. 14.3 w układzie $u_2 = f(v_2)$.



Rys. 14.3 Rozkład obserwacji w przestrzeni $u_2 = f(v_2)$.

15. ANALIZA CZYNNIKOWA I GŁÓWNYCH SKŁADOWYCH

U podstaw analizy czynnikowej leży założenie, że w zespole M cech pierwotnych opisujących daną populację wielowymiarową, ukryte są pewne wspólne czynniki, a w najprostszym przypadku jeden, będący źródłem wspólnej informacji tkwiącej w owych zmiennych pierwotnych. Celem analizy czynnikowej jest wykrycie tych wspólnych czynników.

Zakładamy zatem istnienie macierzy danych określonej jak poniżej:

$$X = \{x_{ik}\} \qquad i = 1, ..., M \quad k = 1, ..., N \tag{15.1}$$

będącej zbiorem N wektorów cech w przestrzeni E^M:

$$X_j = (x_{1j}, x_{2j}, ..., x_{Mj})^T . \qquad j = 1, ... , N \tag{15.2}$$

Zmienne X_j są ze sobą powiązane, a istniejące między nimi zależności można scharakteryzować współczynnikami korelacji:

$$r_{jl} = \frac{s_{jl}^2}{s_j s_l} \qquad j, l = 1, ..., N \tag{15.3}$$

gdzie s_{jl}^2 jest kowariancją między zmiennymi X_j oraz X_l, natomiast s_j, s_l są odchyleniami standardowymi tych właśnie zmiennych.

Wyliczone według wzoru (15.3) współczynniki korelacji tworzą macierz korelacji R

$$R = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1N} \\ r_{21} & r_{22} & \cdots & r_{2N} \\ \cdots & \cdots & \cdots & \cdots \\ r_{N1} & r_{N2} & \cdots & r_{NN} \end{bmatrix} \tag{15.4}$$

w której

$$r_{jj} = 1, \qquad j = 1, ..., N .$$

Macierz korelacji nie ulegnie zmianie, jeśli od zmiennych X_j określonych zależnością (15.2) przejdzie się do zmiennych standaryzowanych:

$$Z_j = (z_{1j}, z_{2j}, ..., z_{Mj})^T , \qquad j = 1, ..., N \tag{15.5}$$

gdzie

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} , \qquad i = 1, ..., M \; j = 1, ..., N$$

oraz

$$\bar{x}_j = \frac{1}{M} \sum_{i=1}^{M} x_{ij} \tag{15.6}$$

jest średnią j-tej cechy.

Użycie zmiennych standaryzowanych lub wyrażonych w jednostkach naturalnych da w efekcie inny układ czynników. Jeśli np. jedna ze zmiennych wyrażonych w jednostkach naturalnych wykazuje w ramach zebranych danych wielokrotnie silniejsze zróżnicowanie niż pozostałe zmienne, to wyznaczony dla tego zbioru danych pierwszy czynnik będzie praktycznie równy tej zmiennej. Standaryzacja zmiennych jest równoważna z założeniem, że wagi jakie przypisujemy poszczególnym zmiennym przy pomiarze stopnia zróżnicowania danego zbioru obserwacji są jednakowe.

Istnienie korelacji między zmiennymi nasuwa przypuszczenie, że istnieją pewne czynniki wpływające na te zmienne. Celem metod analizy czynnikowej i głównych składowych jest wyodrębnienie najistotniejszych czynników spośród zestawu N zmiennych pierwotnych opisujących badane zagadnienie.

Zakłada się zatem, że zmienne pierwotne mają pewne wspólne właściwości. Zmienne te mają również specyficzne, swoiste właściwości. W każdej zmiennej można więc wyróżnić dwa składniki: wspólny i swoisty. Wspólne składniki zmiennych można zastąpić pewnymi wielkościami globalnymi, zwanymi czynnikami wspólnymi, których liczba jest mniejsza od liczby zmiennych pierwotnych. Natomiast czynniki swoiste nie mogą być zastąpione wielkościami syntetycznymi.

Założenia powyższe można sformalizować przyjmując, że każda zmienna Z_j[1] zależy w sposób liniowy od pewnej liczby czynników:

---

1 określona wzorem (15.5)

$$Z_j = w_{j1} F_1 + w_{j2} F_2 + \ldots + w_{jL} F_L + w_j U_j \qquad j = 1, 2, ..., N \tag{15.7}$$

gdzie:

Z_j — j-ta standaryzowana zmienna pierwotna,
F_i — i-ty czynnik wspólny:

$$F_i = (f_{1i}, f_{2i}, ..., f_{Mi})^T , \qquad i = 1, ..., L \tag{15.8}$$

U_j — j-ty czynnik swoisty:

$$U_j = (u_{1j}, u_{2j}, ..., u_{Mj})^T , \qquad j = 1, ..., N \tag{15.9}$$

oraz

f_{ki} — to wartość i-tego wspólnego czynnika w k-tej zmiennej,
u_{kl} — to wartość j-tego czynnika swoistego w k-tej zmiennej.

Zakłada się, że czynniki określone wzorami (15.8) oraz (15.9) są wzajemnie nieskorelowane oraz unormowane. Spełniona jest również nierówność $L \leq N$.

Współczynniki w_{ji} oraz w_j noszą nazwę ładunków; i tak
w_{ji} — to ładunek i-tego czynnika wspólnego występujący w j-tej zmiennej Z_j,
w_j — jest ładunkiem j-tego czynnika swoistego występującego w zmiennej Z_j.

Układ równań (15.7) można przedstawić w zapisie macierzowym postaci:

$$Z = W F + U \tag{15.10}$$

gdzie

$$Z = [Z_1, Z_2, ..., Z_N] ,$$
$$F = [F_1, F_2, ..., F_L] ,$$
$$W = \{w_{ji}\}$$
$$U = [w_1 U_1, w_2 U_2, ..., w_j U_j, ..., w_N U_N]$$

Metoda głównych składowych tym różni się od metody analizy czynnikowej, że zamiast układu równań (15.10) wprowadza się prostszy układ postaci:

$$Z = W F \tag{15.11}$$

W tym układzie równań nie występują elementy odpowiadające czynnikom swoistym. Konsekwencją takiego postawienia problemu jest to, że głównych składowych może być tylko N. Liczba czynników wspólnych i swoistych wynosi natomiast $L + N$. Przy praktycznym wykorzystywaniu tych metod nie ma to jednak większego znaczenia, gdyż

liczbę wydzielonych czynników oraz składowych ogranicza się najczęściej do kilku najważniejszych, a równocześnie takich, które nie stwarzają trudności przy merytorycznej interpretacji. Z tych właśnie przyczyn analiza czynnikowa i analiza głównych składowych rozważane są łącznie.

Rozwiązanie układów równań (15.10) lub (15.11) wymaga wykonania dwóch czynności: wyznaczenia ładunków czynnikowych, a następnie obliczenia wartości czynników lub składowych. Przekształcając w tym celu wzór (15.11) otrzymuje się

$$F = W^{-1}Z$$

Dodatkowo zakłada się, że macierz ładunków czynnikowych jest macierzą ortogonalną, co prowadzi do wzoru:

$$F = W^{T}Z \tag{15.12}$$

Podstawową rolę przy obliczaniu ładunków spełnia macierz korelacji $R = \{r_{kn}\}$, $(k, n = 1, 2, ..., N)$, określona wzorem (15.4). Ponieważ jest to macierz korelacji między zmiennymi standaryzowanymi, zatem

$$r_{kn} = \frac{\operatorname{cov}(Z_k, Z_n)}{s_k s_n} = \frac{1}{M}\sum_{m=1}^{M} z_{mk} z_{mn} \tag{15.13}$$

gdyż $\bar{z}_k = \bar{z}_n = 0$ oraz $s_k = s_n = 1$.

Do powyższego wzoru w miejsce standaryzowanych wartości z_{mk} oraz z_{mn} wstawić można równoważne z nimi liniowe kombinacje czynników — według wzoru (15.7) — a po odpowiednich przekształceniach otrzyma się zależność współczynnika korelacji od ładunków czynnikowych:

$$r_{kn} = w_{k1} w_{n1} + w_{k2} w_{n2} + \ldots + w_{kM} w_{nM} \tag{15.14}$$

Współczynnik korelacji dla dwóch różnych zmiennych równy jest zatem sumie iloczynów ładunków czynnikowych znajdujących się przy wspólnych czynnikach tych zmiennych.

Dowodzi się również, iż współczynnik korelacji między standaryzowaną n-tą zmienną pierwotną a m-tym czynnikiem jest równy odpowiedniemu ładunkowi czynnikowemu, tzn:

$$r_{z_n F_m} = w_{nm} \tag{15.15}$$

Wariancję każdej zmiennej

$$v_n = 1 = \frac{1}{M} \sum_{m=1}^{M} (z_{mn})^2 \tag{15.16}$$

rozłożyć można na dwa składniki

$$v_n = h_n^2 + w_n^2 \tag{15.17}$$

Składnik w_n^2 nazywa się „swoistością", natomiast składnik h_n^2 nosi nazwę zasobu zmienności wspólnej i wyraża się wzorem

$$h_n^2 = w_{n1}^2 + w_{n2}^2 + \ldots + w_{nL}^2 \tag{15.18}$$

Analiza czynnikowa stara się w sposób możliwie pełny wyjaśnić wariancję zmiennych pierwotnych przez czynniki wspólne. Prowadzi to w konsekwencji do eliminacji wpływu czynników swoistych. Porównanie wzorów (15.4), (15.14), (15.17) i (15.18) umożliwia zapis macierzy korelacji R w funkcji czynników wspólnych i swoistych — czynniki swoiste występują jednak jedynie na głównej przekątnej (jedynki zastąpione są tam sumą $h_n^2 + w_n^2$). Eliminowanie niepożądanych czynników swoistych polega na tym, że na głównej przekątnej macierzy korelacji umieszcza się jedynie zasoby zmienności wspólnej.

Macierz korelacji nazywa się wówczas zredukowaną macierzą korelacji i oznacza jako

$$R' = \begin{bmatrix} h_1^2 & r_{12} & \ldots & r_{1N} \\ r_{21} & h_2^2 & \ldots & r_{2N} \\ \ldots & \ldots & \ldots & \ldots \\ r_{N1} & r_{N2} & \ldots & h_N^2 \end{bmatrix} \tag{15.19}$$

Użycie zredukowanej macierzy korelacji upraszcza zagadnienie analizy czynnikowej, podstawowy układ równań ma bowiem teraz postać

$$Z' = W F^T \tag{15.20}$$

Zredukowana macierz korelacji wyraża się wzorem

$$R' = W W \tag{15.21}$$

czyli

$$\begin{bmatrix} h_1^2 & r_{12} & \dots & r_{1N} \\ r_{21} & h_2^2 & \dots & r_{2N} \\ \dots & \dots & \dots & \dots \\ r_{N1} & r_{N2} & \dots & h_N^2 \end{bmatrix} = \begin{bmatrix} w_{11} & w_{12} & \dots & w_{1L} \\ w_{21} & w_{22} & \dots & w_{2L} \\ \dots & \dots & \dots & \dots \\ w_{N1} & w_{N2} & \dots & w_{NL} \end{bmatrix} \begin{bmatrix} w_{11} & w_{21} & \dots & w_{N1} \\ w_{12} & w_{22} & \dots & w_{N2} \\ \dots & \dots & \dots & \dots \\ w_{1L} & w_{2L} & \dots & w_{NL} \end{bmatrix}$$

Jest to podstawowy wzór służący do wyliczania ładunków czynnikowych (sam sposób wyliczania będzie podany dalej). Po ich wyliczeniu korzysta się z wzoru (15.12), tzn.:

$$F = W^T Z$$

aby obliczyć elementy macierzy wartości czynnikowych.

Wprowadzone uproszczenie zmienia niektóre z relacji łączących wariancję i współczynniki korelacji z ładunkami czynnikowymi. Wariancje zmiennych Z' nie są równe wariancji całkowitej, lecz zasobowi zmienności wspólnej. Natomiast współczynnik korelacji dwóch różnych zmiennych nie ulega zmianie, tzn.

$$r'_{kn} = r_{kn} = w_{k1} w_{n1} + w_{k2} w_{n2} + \dots + w_{kM} w_{nM} \tag{15.22}$$

Dla $k = n$ zachodzi natomiast

$$r'_{kk} = w_{k1}^2 + w_{k2}^2 + \dots + w_{kM}^2 = h_k^2$$

Korzystanie ze zredukowanej macierzy korelacji wymaga wstawienia na głównej przekątnej nieznanych zasobów zmienności wspólnej. Wartości te nie są wyznaczane eksperymentalnie, lecz są szacowane przy pomocy różnych metod. Najprostszym i najczęściej stosowanym oszacowaniem jest przyjęcie jako h_n^2 najwyższej wartości ze zbioru współczynników korelacji danej zmiennej z pozostałymi zmiennymi. Wybór metody nie jest tu krytyczny, gdyż wartości h_n^2 otrzymywane różnymi stosowanymi metodami niewiele różnią się między sobą.

Podstawowymi wielkościami w analizie czynnikowej są współczynniki korelacji między wszystkimi parami zmiennych. Interpretacja geometryczna wielkości występujących w analizie czynnikowej pozwala na traktowanie poszczególnych zmiennych jako wektorów. Współczynniki korelacji są wówczas cosinusami kątów między wektorami-zmiennymi. Aby wykazać słuszność tego stwierdzenia skorzystamy ze wzoru na współczynnik korelacji (zmienne są standaryzowane!):

$$r_{kn} = \frac{\sum_{m=1}^{M} z_{mk} z_{mn}}{\sqrt{\sum_{m=1}^{M} z_{mk}^2} \sqrt{\sum_{m=1}^{M} z_{mn}^2}} \tag{15.23}$$

Licznik tego wzoru jest iloczynem skalarnym dwóch wektorów zmiennych Z_k i Z_n:

$$\sum_{m=1}^{M} z_{mk} z_{mn} = |\vec{Z_k}| \, |\vec{Z_n}| \cos \alpha \tag{15.24}$$

a mianownik jest iloczynem długości tych samych wektorów. Wzór (23) przyjmie zatem postać:

$$r_{kn} = \frac{|\vec{Z_k}| \, |\vec{Z_n}| \cos \alpha}{|\vec{Z_k}| \, |\vec{Z_n}|} = \cos \alpha \tag{15.25}$$

Zgodnie z warunkiem (15.22) współczynniki korelacji różnych zmiennych, tzn. elementy spoza głównej przekątnej macierzy R i R 'są sobie równe, tzn.:

$$r_{kn} = r_{kn}' = \cos \alpha \tag{15.26}$$

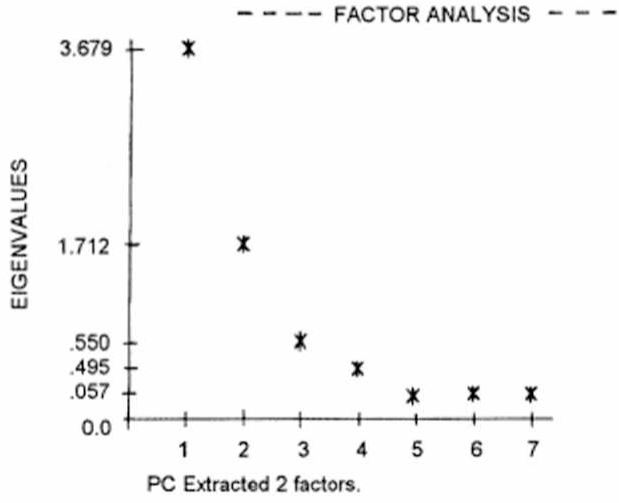
Elementy znajdujące się na głównej przekątnej macierzy R są interpretowane jako długości zmiennych-wektorów. Elementy na głównej przekątnej macierzy R ' są zasobami zmienności wspólnej, a więc obejmują tylko pewną część „długości" wektorów-zmiennych Z_k

$$|\vec{Z_k}| = \sqrt{h_k^2} \leq 1$$

Zatem macierze R i R ' zawierają wszystkie podstawowe informacje o zmiennych. Elementy spoza głównej przekątnej określają kąty między wektorami, a elementy z głównej przekątnej określają długości wektorów. Macierze R i R ' wyznaczają w jednoznaczny sposób całą konfigurację wektorów. Również pozostałe wielkości występujące w analizie czynnikowej możemy zinterpretować geometrycznie. Na podstawie warunku (15.7) interpretujemy czynniki F_i — są one wektorami o jednostkowej długości, a kąty między nimi mają 90°. Czynniki są więc interpretowane jako osie prostokątnego układu współrzędnych.
Ładunki czynnikowe interpretuje się natomiast jako rzuty wektorów-zmiennych na osie-czynniki. Zależność (15.15) wyjaśnia, że ładunek czynnikowy m-tego czynnika i n-tej zmiennej jest współczynnikiem korelacji między n-tą zmienną i m-tym czynnikiem jako cosinus kąta zawartego między tymi wielkościami. A zatem

$$w_{nm} = r_{Z_n F_m} = \cos \beta \tag{15.27}$$

Zgodnie z (15.27) i rysunkiem 15.1 ładunek czynnikowy w_{nm} możemy traktować jako rzut n-tej zmiennej na m-ty czynnik.



Rys. 15.1 Wartości własne.

Graficzna interpretacja podstawowych pojęć analizy czynnikowej ułatwia zrozumienie jej istoty. Jest to wzajemne usytuowanie dwóch zestawów wektorów. Jeden zestaw jest konfiguracją wektorów-zmiennych, która w jednoznaczny sposób jest określona przez macierze korelacji R i R'. Drugi zestaw to wektory-czynniki, które interpretuje się jako osie układu współrzędnych. Na układ osi-czynników jest nałożony układ wektorów-zmiennych. Wektory-zmienne są rzutowane na osie-czynniki. Rzuty te są ładunkami czynnikowymi. Rezultatem tego postępowania jest zastąpienie wektorów-zmiennych wektorami-czynnikami, które są ortogonalne (i dlatego każda z tych wielkości wnosi innego rodzaju informacje) i których jest znacznie mniej niż zmiennych. Mniejsza liczba czynników niż zmiennych bierze się stąd, że wiele ze zmiennych jest silnie skorelowanych, co znaczy, że wnoszą podobne informacje o badanych zjawiskach. Takie silnie skorelowane zmienne są więc zastępowane mniej licznymi zbiorami ortogonalnych czynników.

Korzystając z geometrycznej interpretacji pojęć analizy czynnikowej można również wyjaśnić podstawowy problem tej analizy, a mianowicie brak jednoznacznego rozwiązania tego zagadnienia. Wada ta wynika stąd, że układ odniesienia, tj. układ osi-czynników, nie jest jednoznacznie ustalony. To niejednoznaczne usytuowanie układu odniesienia wpływa oczywiście na wartości ładunków czynnnikowych, ponieważ rzuty wektorów-zmiennych na osie-czynniki zależą od położenia tychże osi-czynników. Tak więc pomimo jednozna-

czności konfiguracji wektorów-zmiennych dochodzi do niejednoznacznego rozwiązania zagadnienia analizy czynnikowej.

Podstawowe równanie analizy czynnikowej (15.21) wykazuje niejednoznaczność wyznaczania macierzy W. Jedna z metod wyznaczania macierzy W polega na wyborze czynników w sposób zmniejszający ich udział w ogólnej zmienności wspólnej, zdefiniowanej jako

$$V = \sum_{n=1}^{N} h_n^2 = \sum_{n=1}^{N} \sum_{l=1}^{L} w_{nl} \tag{15.28}$$

Iteracyjne postępowanie rozpoczyna się od określenia czynnika F_1, którego udział w ogólnej zmienności wspólnej jest z założenia największy. Tworzy się funkcję N zmiennych stanowiącą sumę kwadratów ładunków czynnika F_1 we wszystkich zmiennych Z_n ($n = 1, ..., N$):

$$V_1 = \sum_{n=1}^{N} w_{n1}^2 \tag{15.29}$$

Funkcję tę maksymalizuje się, przy dodatkowych nałożonych na czynniki warunkach określonych wzorem (15.21). Po wyznaczeniu ładunków[2]

$$w_1 = (w_{11}, w_{21}, ..., w_{n1}, ..., w_{N1})$$

tworzy się macierz

$$W_1 = w_1 w_1^T$$

a następnie macierz

$$R_1 = R' - W_1 \tag{15.30}$$

nazywaną macierzą pozostałości korelacyjnej po pierwszym czynniku.

W analogiczny sposób wyznacza się ładunki drugiego czynnika F_2 maksymalizując funkcję

$$V_2 = \sum_{n=1}^{N} w_{n2}^2 \quad ,$$

2 w sposób podany poniżej

przy założeniu, że ładunki czynnikowe spełniają warunki wynikające ze wzoru (15.30). Postępując tak dalej osiąga się żądany poziom wyjaśniania wariancji zmiennych Z_n, np. 75%.

Wyjaśnienia wymaga sposób poszukiwania warunkowego ekstremum funkcji V_i. Stosuje się tu metodę mnożników Lagrange'a, co sprowadza się (dowód pomijamy) do wyznaczenia wszystkich różnych od zera, uporządkowanych nierosnąco wartości własnych macierzy R ' oraz odpowiadających im wektorów własnych. Macierz R ' jest symetryczna, zatem jej wartości własne są rzeczywiste a odpowiadające im wektory własne ortogonalne.

Ładunki czynnikowe oblicza się stosując wzór

$$w_l = \sqrt{\lambda_l}\, A_l \tag{15.31}$$

gdzie

$w_l = (w_{1l}, w_{2l}, ..., w_{Nl})^T$ — jest wektorem ładunków czynnikowych występujących przy l-tym czynniku;

$\lambda = (\lambda_1, \lambda_2, ..., \lambda_L)$ — jest wektorem wartości własnych;

$A_l = (a_{1l}, a_{2l}, ..., a_{Nl})$ — jest unormowanym wektorem własnym odpowiadającym l-tej wartości własnej.

Zatem ładunek czynnikowy n-tej zmiennej i l-tego czynnika jest obliczany jako

$$w_{nl} = \sqrt{\lambda_l}\, \frac{a_{nl}}{\sqrt{\sum_{n=1}^{N} a_{nl}^2}} \tag{15.32}$$

Wyznaczanie ładunków w metodzie głównych składowych różni się tylko tym, że macierzą wyjściową jest macierz korelacji. W macierzy tej znajduje się wartości własne oraz wektory własne i na ich podstawie ze wzoru (15.32) wyznacza się wielkości w_{nl}.

Zauważmy jeszcze, że jeśli D jest macierzą ortogonalną, to transformacja czynników

$$W' = W\, D$$

nie zmienia struktury macierzy korelacji określonej wzorem

$$R' = W\, W^T\ ,$$

ponieważ:

$$W\, D\ (W\, D)^T = W\, D\, D^T W^T = W\, W^T\ .$$

Opisanemu przekształceniu geometrycznie odpowiada obrót kierunków głównych określających składowe główne. Obrót taki pozwala łatwiej zinterpretować otrzymane wyniki. Normalnie wszystkie ładunki mają wartości z przedziału (–1, 1) i na ich podstawie trudno jest określić najważniejsze czynniki. Po dokonaniu ortogonalnej rotacji, elementy macierzy ładunków W dają się łatwo interpretować. Współczynniki znaczące mają po rotacji wartości bliskie jedynce, natomiast mało znaczące są bliskie zeru. Obrotu dokonuje się zatem tak, aby ładunki przy cechach maksymalnie się różnicowały, wykorzystując warunek:

$$\text{var } w = \sum_{n,m} (w_{nm} - \overline{w}_n)^2 = \max \ .$$

Stosowna procedura realizująca opisaną rotację ortogonalną nosi nazwę metody varimax.

Przykład.

Korzystamy z danych zebranych w dodatku 2, a dotyczących nadczynności gruczołów tarczycowych. Wybieramy pierwszą klasę pacjentów, tzn. tych dla których leczenie zakończyło się pomyślnie. Liczba pacjentów wynosi $M = 16$. Uwzględnimy $N = 7$ cech, a mianowicie $x1$, $x2$, $x3$, $x4$, $x5$, $x6$ oraz $x10$. Wybierzemy metodę głównych składowych i postaramy się znaleźć tyle czynników, by wyjaśniały one co najmniej 75% wariancji zmiennych wyjściowych.

Obliczenia rozpoczynamy od wyliczenia macierzy korelacji, jest ona następująca:

$$R = \begin{bmatrix} 1{,}000 & & & & & & \\ 0{,}683 & 1{,}000 & & & & & \\ 0{,}564 & 0{,}850 & 1{,}000 & & & & \\ 0{,}345 & 0{,}589 & 0{,}811 & 1{,}000 & & & \\ 0{,}222 & 0{,}032 & -{,}223 & -{,}274 & 1{,}000 & & \\ 0{,}559 & 0{,}465 & 0{,}436 & 0{,}355 & 0{,}236 & 1{,}000 & \\ 0{,}533 & 0{,}509 & 0{,}302 & 0{,}299 & 0525 & 0{,}609 & 1{,}00 \end{bmatrix}$$

Rozwiązując stosowne zagadnienie własne otrzymujemy:

Czynnik	Wart. własna	% wariancji	Skum. % warian.
1	3,67944	52,6	52,6
2	1,71232	24,5	77,0
3	0,55037	7,9	84,9
4	0,49498	7,1	92,0
5	0,26060	3,7	95,7
6	0,24574	3,5	99,2
7	0,05654	0,8	100,0

Wyliczone wartości własne pokazane są na rysunku 15.1. Z przedstawionej powyżej tablicy wynika, że jako zmienne syntetyczne powinniśmy wybrać dwa pierwsze czynniki wyjaśniające łącznie 77% wariancji. Z problemu rozpatrywanego w przestrzeni siedmiowymiarowej potrafimy zatem przejść do przestrzeni dwuwymiarowej (tracąc niewiele informacji) i przeliczając zmienne według wzorów:

$$F_1 = 0{,}797x1 + 0{,}889x_2 + 0{,}851x_3 + 0{,}715x_4 + 0{,}132x_5 + 0{,}723x6 + 0{,}693x10 \ ,$$

$$F_2 = 0{,}190x1 - 0{,}142x2 - 0{,}453x3 - 0{,}503x4 + 0{,}895x_5 + 0{,}299x6 + 0{,}554x10 \ .$$

Wartości stałe występujące w tych wzorach składają się na macierz W^T, tzn. transponowaną macierz ładunków czynnikowych.

DODATEK 1.
ZMIENNE LOSOWE I ICH ROZKŁADY

Zmienne losowe

Z wystarczającą dla potrzeb tego skryptu dokładnością można przyjąć, że zmienna losowa jest to taka zmienna, która w wyniku doświadczenia przyjmuje jedną i tylko jedną możliwą wartość, przy czym wartości tej nie można z góry przewidzieć, gdyż zależy ona od przyczyn losowych. W biometrii używa się również często pojęcia *cechy statystycznej* — jest to odpowiednik zmiennej losowej.

Jeśli zbiór wartości jakie może przyjmować zmienna losowa jest skończony lub przeliczalny, to mamy do czynienia ze *zmienną dyskretną* (inna nazwa: *zmienna skokowa*). Jeśli natomiast zbiór wartości zmiennej losowej jest nieprzeliczalny to nosi ona nazwę *zmiennej losowej ciągłej*.

Rozkład zmiennej losowej

Ze zmienną losową związane jest pojęcie rozkładu prawdopodobieństwa tej zmiennej (krócej: rozkładu zmiennej) oraz jej dystrybuanty. W przypadku zmiennej losowej dyskretnej X rozkładem tej zmiennej nazywa się zestawienie wszystkich przybieranych przez nią wartości wraz z odpowiadającymi im prawdopodobieństwami. Innymi słowy, jeśli przez p_k oznaczy się prawdopodobieństwo, że zmienna losowa X przyjmie wartość x_k, czyli:

$$P(X = x_k) = p_k$$

to wówczas rozkładem zmiennej losowej X jest zbiór wszystkich par

$$(x_k, p_k) \ .$$

Prawdopodobieństwa p_k spełniają oczywisty warunek:

$$\sum_k p_k = 1$$

Rozkład zmiennej losowej dyskretnej może być zadany za pomocą tabeli, analitycznie lub graficznie.

Znając rozkład zmiennej losowej dyskretnej X możemy zdefiniować funkcję $F(x)$, która dla każdej wartości x określa prawdopodobieństwo zdarzenia, że zmienna losowa X przyjmie wartość mniejszą od x, tzn.:

$$F(x) = P(X < x)$$

Funkcja ta nazywa się *dystrybuantą zmiennej losowej*. Dystrybuanta dowolnej zmiennej losowej jest funkcją niemalejącą przyjmującą wartości z przedziału <0, 1>. Dla zmiennej losowej skokowej dystrybuanta przedstawiona jest wzorem

$$F(x) = \sum_{x_i < x} P\,(X = x_i) = \sum_{x_i < x} P_i$$

gdzie symbol $x_i < x$ pod znakiem sumy oznacza sumowanie rozciągnięte na te wszystkie wartości x_i, które są mniejsze od x.

Prawdopodobieństwo, że zmienna losowa X (dyskretna lub ciągła) przyjmie wartości należące do przedziału $<a, b>$ można wyrazić za pomocą jej dystrybuanty, prawdopodobieństwo to jest równe przyrostowi dystrybuanty w tym przedziale, czyli:

$$P\,(a \le X < b) = F\,(b) - F\,(a)$$

Jeżeli mamy do czynienia ze zmienną losową ciągłą to nie potrafimy określić jej rozkładu wyliczając wszystkie jej możliwe wartości wraz z odpowiadającymi im prawdopodobieństwami. Możemy jednak posłużyć się dystrybuantą tej zmiennej losowej oraz zdefiniować funkcję gęstości prawdopodobieństwa jako pochodną dystrybuanty, tzn.:

$$f\,(x) = F\,'(x)$$

czyli

$$F\,(x) = \int_{-\infty}^{x} f\,(t)\,dt$$

Funkcja gęstości prawdopodobieństwa danej zmiennej losowej X jest zatem miarą prawdopodobieństwa zajścia następującego zdarzenia:

$$(x \leq X \leq x + dx)$$

W problemach spotykanych w praktyce zmienne losowe ciągłe posiadają w każdym punkcie przedziału określoności (za wyjątkiem co najwyżej skończonej liczby punktów na każdym skończonym odcinku) ciągłą gęstość prawdopodobieństwa.

Podstawowe własności funkcji gęstości prawdopodobienstwa to:

$$f(x) \geq 0$$

$$P(X < a) = F(a) = \int_{-\infty}^{a} f(x)\, dx$$

$$P(a \leq X < b) = \int_{a}^{b} f(x)\, dx = F(b) - F(a)$$

a w szczególności

$$\int_{-\infty}^{\infty} f(x)\, dx = 1$$

Wykres funkcji gęstości prawdopodobieństwa nosi nazwę *krzywej rozkładu.*

Parametry rozkładu zmiennej losowej

Rozkład prawdopodobieństwa w przypadku zmiennej losowej dyskretnej (albo funkcja gęstości prawdopodobieństwa w przypadku zmiennej losowej ciągłej) lub dystrybuanta charakteryzują w pełni zmienną losową. W praktyce bardzo często wystarczy posługiwać się jedynie pewnymi wartościami opisującymi tę zmienną. Najczęściej używane charakterystyki liczbowe rozkładu prawdopodobieństwa zmiennej losowej to:

— wartość oczekiwana,
— wariancja,
— odchylenie standardowe.

Oprócz nich definiuje się również takie charakterystyki, jak: modalną, odchylenie przeciętne, współczynnik zmienności czy momenty — nie będziemy ich tu jednak omawiać.

Wartość oczekiwana zmiennej losowej dyskretnej X definiowana jest jako:

$$E(X) = \sum_{k=1}^{N} x_k p_k$$

natomiast dla zmiennej losowej ciągłej

$$E(X) = \int_{-\infty}^{\infty} xf(x)\,dx$$

Wartość oczekiwana określa średnią wartość zmiennej losowej. W jej obliczaniu biorą udział wszystkie wartości zmiennej losowej, jest zatem najbardziej syntetyczną charakterystyką rozkładu zmiennej.

Na ogół oprócz wartości oczekiwanej oblicza się również wariancję zmiennej losowej będącą miarą rozrzutu (rozproszenia) wszystkich wartości zmiennej losowej wokół jej wartości oczekiwanej. Wariancję zmiennej losowej definiuje się jako

$$\sigma^2 = V(X) = E\{[X - E(X)]^2\}$$

czyli

$$V(X) = \sum_{k=1}^{N} [x_k - E(X)]^2 p_k$$

dla zmiennej dyskretnej, oraz

$$V(X) = \int_{-\infty}^{\infty} [x - E(X)]^2 f(x)\,dx$$

dla zmiennej ciągłej.

Natomiast *odchylenie standardowe* definiuje się jako pierwiastek z wariancji (wyraża się ono wówczas w tych samych jednostkach co dana zmienna):

$$\sigma = \sqrt{\sigma^2} = \sqrt{V(X)}$$

Znajomość podstawowych parametrów rozkładów zmiennych losowych, takich jak wartość oczekiwana czy też wariancja pozwala na wyznaczanie tych parametrów dla funkcji owych zmiennych losowych (np. ich kombinacji liniowej). Korzysta się wówczas z podstawowych własności wartości oczekiwanej i wariancji przedstawionych poniżej.

W1. Wartość oczekiwana wielkości stałej jest równa tej stałej, tzn.:

$$E(C) = C \qquad \text{gdzie } C = \text{const}$$

W2. Wartość oczekiwana iloczynu danej zmiennej losowej X i stałej C jest równa iloczynowi tej stałej i wartości oczekiwanej zmiennej X, tzn.:

$$E(C * X) = C * E(X) \qquad \text{gdzie } C = \text{const}$$

W3. Wartość oczekiwana sumy dwóch (ogólnie: dowolnej skończonej liczby) zmiennych losowych X oraz Y równa się sumie wartości oczekiwanych tych zmiennych, tzn.:

$$E(X + Y) = E(X) + E(Y)$$

W4. Wartość oczekiwana różnicy dwóch zmiennych losowych X oraz Y jest równa różnicy wartości oczekiwanych tych zmiennych, tzn.:

$$E(X - Y) = E(X) - E(Y)$$

W5. Wartość oczekiwana iloczynu dwóch (ogólnie: dowolnej skończonej liczby) niezależnych[1] zmiennych losowych X oraz Y równa się iloczynowi wartości oczekiwanych tych zmiennych losowych, tzn.:

$$E(X * Y) = E(X) * E(Y) \qquad \text{o ile } X \text{ i } Y \text{ są niezależne.}$$

W6. Wariancja wielkości stałej jest równa zeru, tzn.:

$$V(C) = 0 \qquad \text{gdzie } C = \text{const}$$

W7. Wariancja iloczynu zmiennej losowej i wartości stałej jest równa iloczynowi kwadratu tej stałej i wariancji zmiennej losowej, tzn.:

$$V(C * X) = C^2 * V(X) \qquad \text{gdzie } C = \text{const}$$

W8. Wariancja sumy dwóch (w ogólnym przypadku: dowolnej skończonej liczby) niezależnych zmiennych losowych X oraz Y równa się sumie wariancji tych zmiennych[2], tzn.:

$$V(X + Y) = V(X) + V(Y), \qquad \text{o ile } X \text{ i } Y \text{ są niezależne}$$

W9. Wariancja różnicy dwóch niezależnych zmiennych losowych X oraz Y jest równa sumie wariancji tych zmiennych losowych, tzn.:

---

1 Nieco uprzedzając rozważania dotyczące zmiennych dwuwymiarowych można stwierdzić, że dwie zmienne losowe X oraz Y są niezależne, jeżeli dystrybuanta zmiennej losowej dwuwymiarowej (X, Y) jest równa iloczynowi dystrybuant składowych.

2 Jeżeli zmienne są zależne, to po prawej stronie wzoru pojawia się dodatkowa składowa zwana kowariancją.

$$V(X - Y) = V(X) + V(Y), \qquad \text{o ile } X \text{ i } Y \text{ są niezależne}$$

Zmienna losowa standaryzowana

Rozpatrzmy dowolną zmienną losową X, dla której określiliśmy zarówno wartość oczekiwaną $E(X)$ jak też wariancję $V(X)$.
Wprowadzamy teraz nową zmienną losową U zdefiniowaną poniżej:

$$U = \frac{X - E(X)}{\sqrt{V(X)}}, \qquad V(X) > 0$$

Zmienna U ma natępujące parametry[3]

$$E(U) = 0, \qquad V(U) = 1$$

Taki sposób uniezależnienia się od konkretnych wartości parametrów zmiennej losowej (a zatem sprowadzenia całej rodziny rozkładów zmiennej losowej do jednego rozkładu) nazywa się jej *standaryzacją*. Standaryzacja jest najczęściej wykorzystywana przy odczycie z tablic statystycznych.

Podstawowe rozkłady teoretyczne

Jest rzeczą zastanawiającą, że olbrzymia większość rozpatrywanych w praktyce cech statystycznych podlega jednemu z kilku uniwersalnych rozkładów teoretycznych. Trzema głównymi rozkładami, które omówimy poniżej są: rozkład dwumianowy, rozkład Poissona oraz rozkład normalny.

Dodatkowo przedstawimy również występujące w testowaniu statystycznym rozkłady: χ^2 (chi — kwadrat), t Studenta oraz F Snedecora. Znajomość tych rozkładów nie jest niezbędna podczas testowania hipotez statystycznych, pozwala jednak na zrozumienie sposobu konstrukcji testów statystycznych.

Stablicowane wartości wszystkich wspomnianych rozkładów znajdują się na końcu skryptu.

---

3 Wyprowadzenie tych zależności proponujemy jako ćwiczenie ugruntowywujące zrozumienie własności **W1 – W9**

Rozkład dwumianowy

Bardzo często mamy do czynienia z doświadczeniem, którego wyniki przybierają zawsze jedną z dwu wzajemnie wykluczających się wartości a prawdopodobieństwa uzyskania tych wyników pozostają te same przez cały czas prób. Zazwyczaj oznaczamy te prawdopodobieństwa przez p oraz q i wynik o prawdopodobieństwie p nazywamy sukcesem, natomiast wynik o prawdopodobieństwie q — porażką.
Zachodzi oczywiście:

$$p + q = 1 \qquad \text{czyli} \qquad q = 1 - p$$

Załóżmy teraz, że przeprowadzamy całą serię takich doświadczeń (ściślej mówiąc powtarzamy doświadczenie n razy). Ponieważ doświadczenia są niezależne, zatem prawdopodobieństwo uzyskania w serii n doświadczeń k pierwszych sukcesów wynosi

$$p^k q^{(n-k)}$$

Jeśli interesuje nas tylko ogólna liczba sukcesów osiągniętych w wyniku n prób, a nie ich kolejność to prawdopodobieństwo uzyskania k sukcesów w n próbach jest równe

$$\binom{n}{k} p^k q^{(n-k)} = \frac{n!}{k!\,(n-k)!}\, p^k q^{(n-k)}$$

Powstały w ten sposób rozkład prawdopodobieństwa nazywa się rozkładem dwumianowym. Zachodzi zatem:

$$P(X^n = k) = \frac{n!}{k!\,(n-k)!}\, p^k q^{(n-k)}$$

gdzie X^n oznacza zmienną losową składającą się z n prób. Rozkład ten zależy zatem od dwóch wielkości: prawdopodobieństwa sukcesu — p, oraz wielkości serii doświadczeń — n.

Rozkład dwumianowy (zwany inaczej rozkładem Bernoulliego) często występuje w praktyce. Przykładowo[4] przypuśćmy, że dla n ludzi zmierzono ciśnienie krwi przed

---

4 Przykład zaczerpnięty z: Feller W.: Wstęp do rachunku prawdopodobieństwa, PWN, Warszawa, 1977

i po podaniu pewnego leku. Oznaczmy wyniki przez $x_1, \ldots, x_n$ oraz $x_1', \ldots, x_n'$. Powiemy, że i-te doświadczenie dało w wyniku sukces, jeżeli $x_i < x_i'$ i porażkę, jeżeli $x_i > x_i'$ (dla uproszczenia załóżmy, że żadne dwa pomiary nie dały tego samego wyniku). Jeżeli lek nie wpływa na ciśnienie krwi, to nasze obserwacje powinny być zgodne z rozkładem dwumianowym z prawdopodobieństwem sukcesu $p = 0{,}5$, natomiast znaczna ilość sukcesów powinna być uznana za wskazówkę, że badany lek ma wpływ na ciśnienie krwi.

Niech X_k będzie ilością sukcesów zanotowaną przy k-tej próbie. Ta zmienna losowa przybiera tylko wartości 0 i 1 z odpowiednimi prawdopodobieństwami p oraz q. Zatem

$$E(X_k) = 0 * q + 1 * p = p,$$

a ponieważ zmienna losowa dwumianowa X_k^n jest sumą n takich niezależnych zmiennych X_k zatem jej wartość oczekiwana wynosi

$$E(X_k^n) = np$$

Analogicznie

$$V(X_k) = E[(X_k - p)^2] = (1-p)^2 p + (0-p)^2 q = pq$$

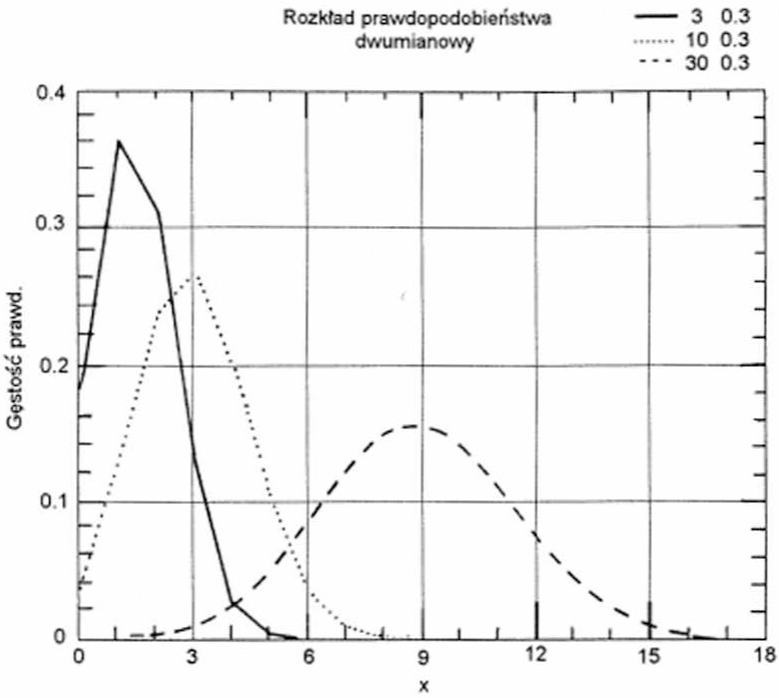
oraz

$$V(X_k^n) = npq$$

czyli

$$\sigma = \sqrt{npq}$$

Rysunek D1.1 przedstawia wykresy rozkładu dwumianowego dla ustalonej wartości prawdopodobieństwa $p = 0{,}3$, i zmieniającej się wartości $n = 5, 10, 30$. Analizując rysunek można zauważyć że wraz ze wzrostem wartości n wykres rozkładu przybiera coraz bardziej symetryczną postać.



Rys. D1.1 Przykład rozkładu dwumianowego

Rozkład Poissona

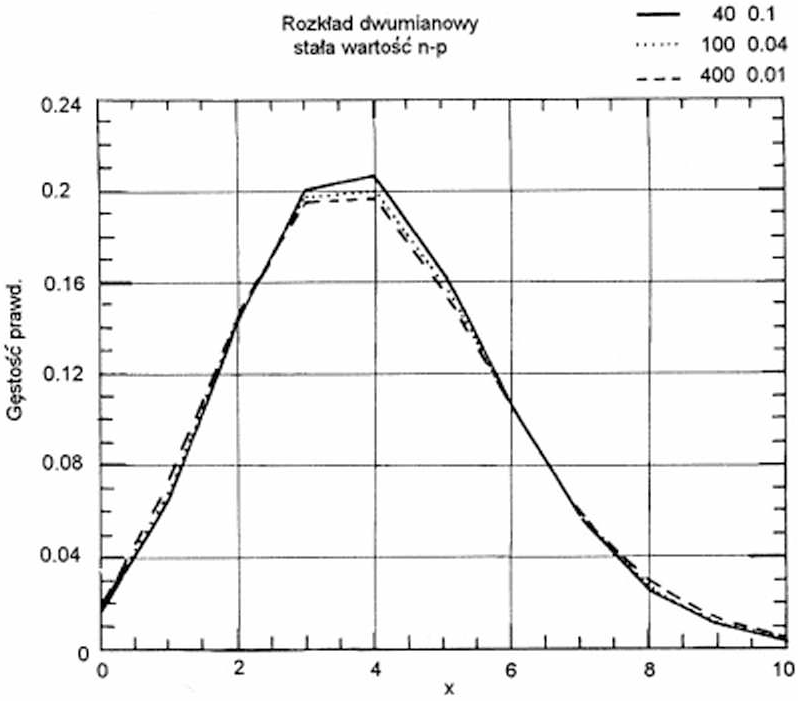
Rozpatrzmy pewien szczególny przypadek rozkładu dwumianowego, dla którego iloczyn $n * p$ ma wartość stałą. Jak widać na rysunku D1.2 wykresy takich rozkładów są bardzo zbliżone. Jeśli założymy ponadto, iż liczba doświadczeń n rośnie nieograniczenie, natomiast prawdopodobieństwo p jest bardzo małe, tzn:

$$n \to \infty$$
$$p \to 0$$
$$n * p = \lambda = \text{const} ,$$

to otrzymamy graniczny przypadek rozkładu dwumianowego zwany rozkładem Poissona określony wzorem

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

gdzie k jest (jak poprzednio) ilością sukcesów.



Rys. D1.2 Rozkład dwumianowy dla $n * p = \text{const.}$

W praktyce im większa jest wartość n tym lepiej rozkład dwumianowy jest zbieżny do rozkładu Poissona. Ilustruje to poniższa tabela.

k	Wartości funkcji prawdopodobieństwa rozkładu dwumianowego dla $n * p = 3$						Wartości f. prawd. rozkładu Poissona
	$p = 0{,}5$ $n = 6$	$p = 0{,}2$ $n = 15$	$p = 0{,}1$ $n = 30$	$p = 0{,}05$ $n = 60$	$p = 0{,}02$ $n = 150$	$p = 0{,}01$ $n = 300$	
0	0,0156	0,0352	0,0424	0,0461	0,0483	0,0490	0,0498
1	0,0937	0,1319	0,1413	0,1455	0,1478	0,1486	0,1494
2	0,2344	0,2309	0,2276	0,2259	0,2248	0,2244	0,2240
3	0,3125	0,2501	0,2361	0,2298	0,2263	0,2252	0,2240
4	0,2344	0,1876	0,1771	0,1724	0,1697	0,1689	0,1680
5	0,0937	0,1032	0,1023	0,1016	0,1011	0,1010	0,1008
6	0,156	0,0430	0,0474	0,0490	0,0499	0,0501	0,0504
7		0,0138	0,0180	0,0199	0,0209	0,0213	0,0216
8		0,0035	0,0058	0,0069	0,0076	0,0079	0,0081
9		0,0007	0,0016	0,0021	0,0025	0,0026	0,0027
10		0,0001	0,0004	0,0006	0,0007	0,0008	0,0008
11			0,0001	0,0001	0,0002	0,0002	0,0002

Rozkład Poissona zależy od jednego parametru λ. Można wykazać, że

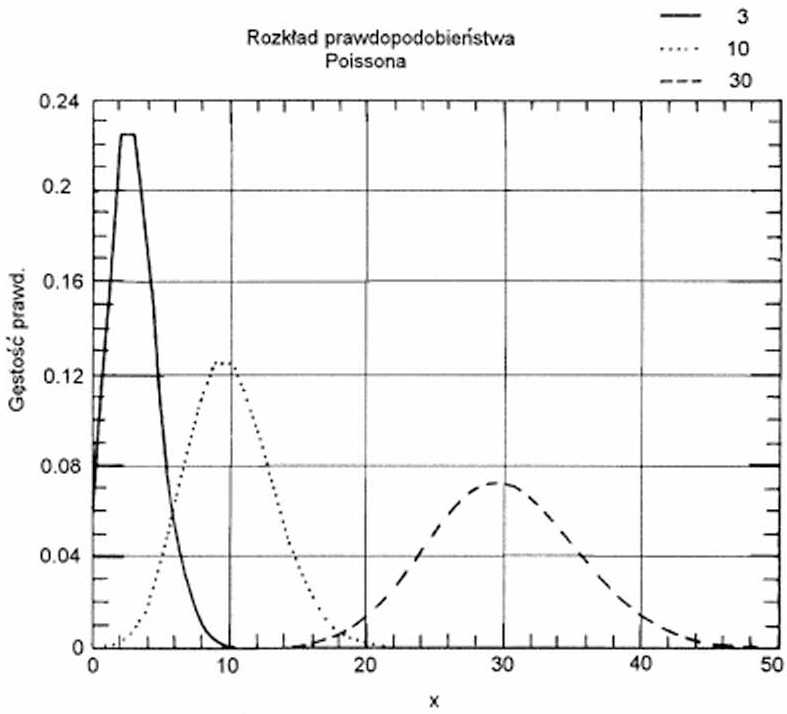
$$E(X) = \lambda$$

oraz

$$V(X) = \lambda$$

Przykłady funkcji gęstości prawdopodobieństwa rozkładu Poissona przedstawione są na rysunku D1.3, dla różnych wartości parametru $\lambda(\lambda = 3, 10, 30)$. Im większa wartość λ tym bardziej symetryczny staje się wykres rozkładu, tym bardziej zbliża się on do rozkładu normalnego.

Ponieważ rozkład Poissona otrzymany został z rozkładu dwumianowego dla dużych wartości n i stałej wartości $\lambda = np$, tzn. małej wartości p, zatem spodziewać się należy zastosowań tego rozkładu do procesów, w których mamy dużą liczbę zdarzeń, ale jedynie niewielki ich ułamek posiada interesującą nas własność. Tak właśnie jest w rzeczywistości. Pyłki w miodzie liczone pod mikroskopem, kolonie bakterii na kwadratach płytki Petriego, gwiazdy w przestrzeni, rodzynki w cieście, nasiona chwastów wśród nasion trawy, skazy w materiale są rozmieszczone zgodnie z prawem Poissona.



Rys. D1.3 Rozkład Poissona.

A oto inny jeszcze przykład[5]. Przy urodzeniu każdy poszczególny człowiek ma małe szanse dożycia do stu lat, ale w dużym społeczeństwie liczba urodzeń w ciągu roku jest wielka. Wskutek wojen, epidemii itp. życia rozmaitych ludzi nie są statystycznie niezależne, ale w pierwszym przybliżeniu możemy porównywać n urodzeń z n próbami Bernoulliego, w których śmierć po stu latach uważamy za sukces. W społeczeństwie ustabilizowanym, gdzie ani wielkość społeczeństwa, ani śmiertelność jego członków nie zmieniają się prędko, rozsądne jest przypuszczenie, że częstość lat, w których umiera dokładnie k stuletnich starców podlega rozkładowi Poissona, przy czym parametr λ zależy od wielkości i stanu zdrowotnego społeczeństwa. Dane statystyczne Szwajcarii potwierdzają to przypuszczenie.

5 Zaczerpnięty z: Feller W.: Wstęp do rachunku prawdopodobieństwa, PWN, Warszawa, 1977

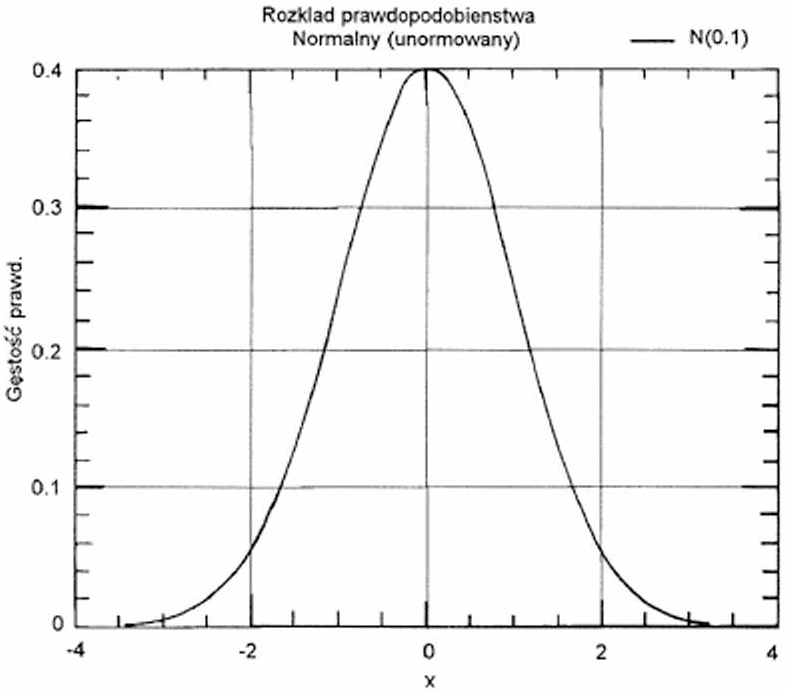
Rozkład normalny

Rozkład normalny zajmuje centralną pozycję pomiędzy licznymi rozkładami ciągłymi stosowanymi w statystyce. Nazwa rozkładu jest nieco myląca, gdyż ma ona znaczenie tylko opisowe, a nie definiujące. Innymi słowy, rozkładów które nie są normalne nie należy uważać za nienormalne. Wiele rozkładów występujących w praktyce zbliżonych jest do rozkładu normalnego, jednak żaden rozkład faktycznych pomiarów nie jest identyczny z rozkładem normalnym, gdyż rozkład skończonej ilości pomiarów może jedynie zbliżyć się do funkcji ciągłej rozkładu normalnego.

Gęstość prawdopodobieństwa rozkładu normalnego opisana jest funkcją

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

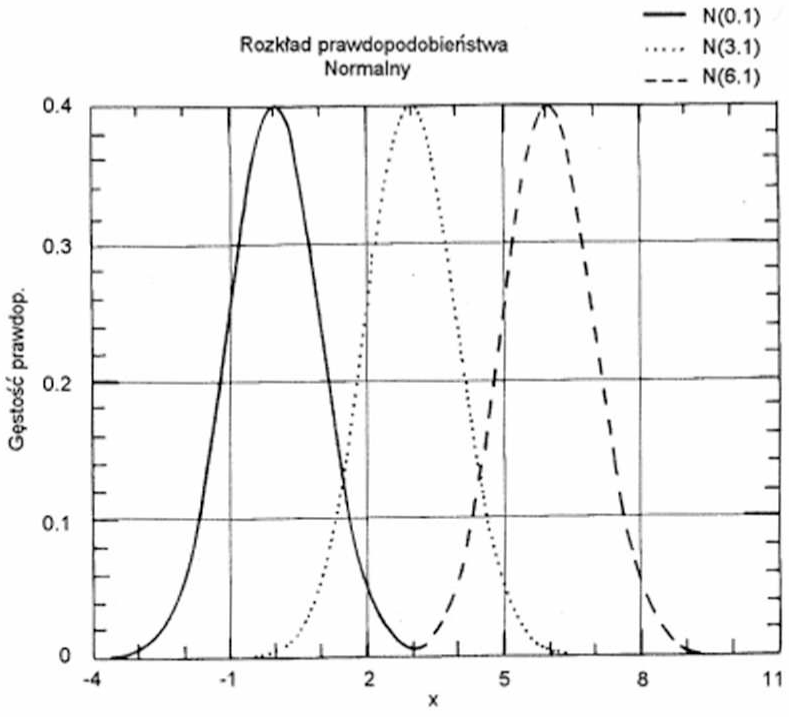
przedstawioną graficznie na rysunku D1.4. Dwa parametry: μ — wartość oczekiwana oraz σ — odchylenie standardowe całkowicie określają kształt krzywej normalnej. Gdy zmienna losowa X ma rozkład normalny o parametrach μ oraz σ, to zapisuje się to jako



Rys. D1.4 Gęstość prawdopodobieństwa rozkładu normalnego.

$x \rightarrow N(\mu, \sigma)$

Na rysunku D1.5 przedstawiono trzy krzywe normalne różniące się wartością parametru μ, a na rysunku D1.6 — różniące się wartością parametru σ. Wartość μ nie wpływa na kształt krzywej, powoduje jej przesunięcie wzdłuż osi odciętych, natomiast zmiana parametru σ powoduje zmianę kształtu krzywej — im mniejsza wartość σ tym smuklejsza krzywa, tym bardziej masa rozkładu skupiona jest wokół średniej.

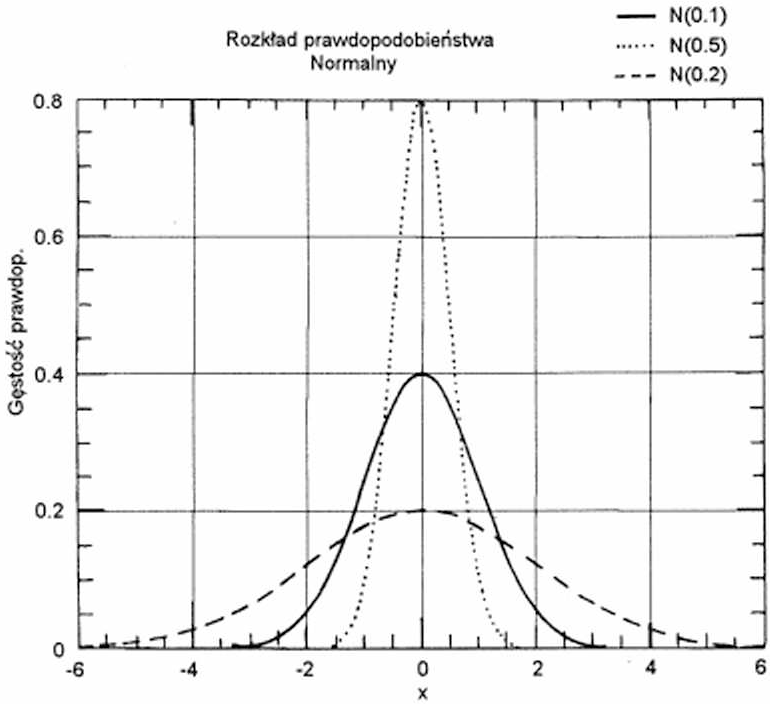


Rys. D1.5 Krzywe rozkładu normalnego.

Gdy zmienna losowa X podlega rozkładowi normalnemu $N(\mu, \sigma)$, wówczas zmienna losowa standaryzowana

$$U = \frac{X - \mu}{\sigma}$$

ma rozkład o parametrach $N(0, 1)$. Rozkład ten (zarówno funkcja gęstości jak i dystrybuanta) został stablicowany, a dzięki standaryzacji z tablic tych można korzystać dla dowolnego rozkładu $N(\mu, \sigma)$ gdyż:



Rys. D1.6 Krzywe rozkładu normalnego.

$$f(x) = \frac{1}{\sigma}\,\phi(u)\,, \qquad F(x) = \Phi(u)$$

gdzie $\phi(u)$ oraz $\Phi(u)$ są odpowiednio funkcją gęstości i dystrybuantą rozkładu $N(0, 1)$. W celu zmniejszenia objętości tablic podawane są wartości jedynie dla nieujemnych wartości $u(u \geq 0)$, natomiast dla wartości ujemnych korzysta się z zależności:

$$\phi(-u) = \phi(u)\,, \qquad \Phi(-u) = 1 - \Phi(u)$$

Pole powierzchni zawartej między krzywą gęstości a osią odciętych jest równe 1, gdyż prawdopodobieństwo, iż zmienna losowa przyjmie wartość w przedziale od $-\infty$ do ∞ jest równe jedności.

Przyjmijmy, że zmienna losowa X ma rozkład $N(\mu, \sigma)$. Obliczymy prawdopodobieństwa, że zmienna X przyjmuje wartość z przedziałów $(\mu - \sigma, \mu + \sigma)$, $(\mu - 2\sigma, \mu + 2\sigma)$, $(\mu - 3\sigma, \mu + 3\sigma)$. Są one oczywiście równe prawdopodobieństwom, że powstała ze

zmiennej X standaryzowana zmienna losowa U przyjmuje wartości z przedziałów, odpowiednio, (–1, 1), (–2, 2), (–3, 3)[6], a zatem

$$P(-1 < U < 1) = \Phi(1) - \Phi(-1) = \Phi(1) - (1 - \Phi(1)) = 2\Phi(1) - 1 = 0{,}6826\,;$$

$$P(-2 < U < 2) = 2\Phi(2) - 1 = 0{,}9545\,;$$

$$P(-3 < U < 3) = 2\Phi(3) - 1 = 0{,}9973\,;$$

Ostatni wzór jest podstawą szeroko stosowanej tzw. *reguły trzech sigm*: *jeżeli zmienna losowa ma rozkład normalny, to wartość bezwzględna odchylenia tej zmiennej od wartości oczekiwanej nie powinna przekraczać potrojonego odchylenia standardowego.*

Prawdopodobieństwo przekroczenia tej wartości wynosi bowiem jedynie 1 – 0,9973 = = 0,0027 = 0,27% !

Dominujące znaczenie rozkładu normalnego wywołane jest przede wszystkim następującymi czynnikami:

— większość cech populacji biologicznych ma rozkład w przybliżeniu normalny, lub rozkład, który można zmienić na normalny po odpowiednim przekształceniu,

— wiele rozkładów w określonych warunkach zbliża się do rozkładu normalnego, przykładem może być tu chociażby rozkład dwumianowy,

— jeżeli zmienna losowa X jest sumą bardzo dużej liczby wzajemnie niezależnych zmiennych losowych, przy czym wpływ każdej z tych zmiennych na sumę jest znikomy, to rozkład zmiennej X jest w przybliżeniu normalny[7]; przybliżenie jest tym lepsze, im liczba zmiennych jest większa.

Rozkład chi-kwadrat (χ^2)

Niech zmienna losowa X ma rozkład prawdopodobieństwa $N(\mu, \sigma)$. Wówczas zmienna standaryzowana $U = (X - \mu)/\sigma$ podlega rozkładowi $N(0, 1)$. Kwadrat zmiennej standaryzowanej normalnej U nosi nazwę zmiennej losowej χ^2 (chi-kwadrat) z jednym stopniem swobody; zatem

---

6 Inaczej mówiąc, są to prawdopodobieństwa tego, że wartość zmiennej losowej X będzie odbiegała od średniej mniej niż, odpowiednio, odchylenie standardowe, dwa odchylenia standardowe, trzy odchylenia standardowe

7 Jest to wniosek wynikający ze znanego z rachunku prawdopodobieństwa centralnego twierdzenia granicznego

$$\chi^2 = \left(\frac{X - \mu}{\sigma}\right)^2 = U^2$$

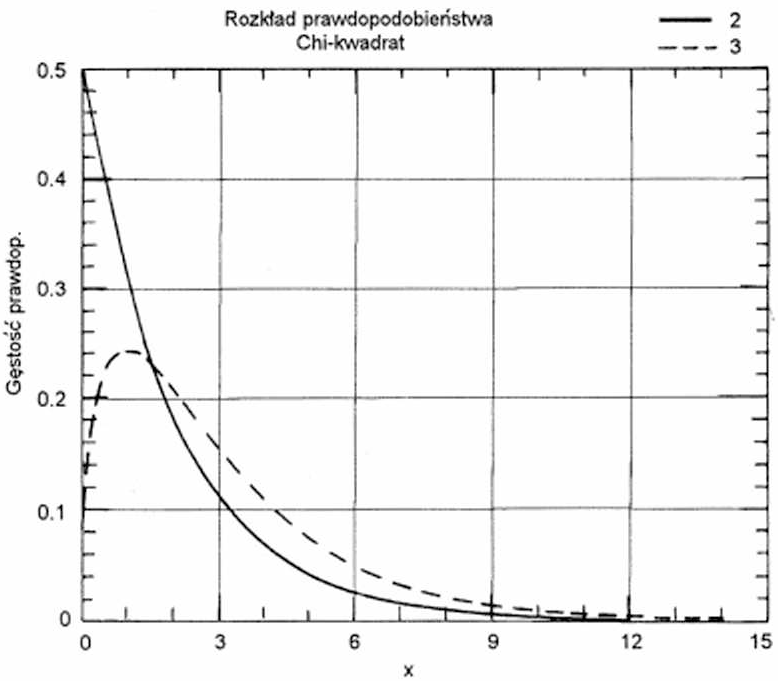
Powyższe wyrażenie obejmuje jeden składnik i dlatego odpowiadająca mu ilość stopni swobody wynosi 1. Jeżeli jednak rozpatrywać będziemy n zmiennych losowych X_i (i = 1, 2, ..., n) o rozkładach odpowiednio $N(\mu_i, \sigma_i)$, to suma kwadratów odpowiadających im zmiennych standaryzowanych U_i, tzn.:

$$\chi^2 = \sum_{i=1}^{n} u_i^2 = \sum_{i=1}^{n} \left(\frac{X_i - \mu_i}{\sigma_i}\right)^2$$

nosi nazwę zmiennej losowej χ^2 z $\nu = n$ stopniami swobody.

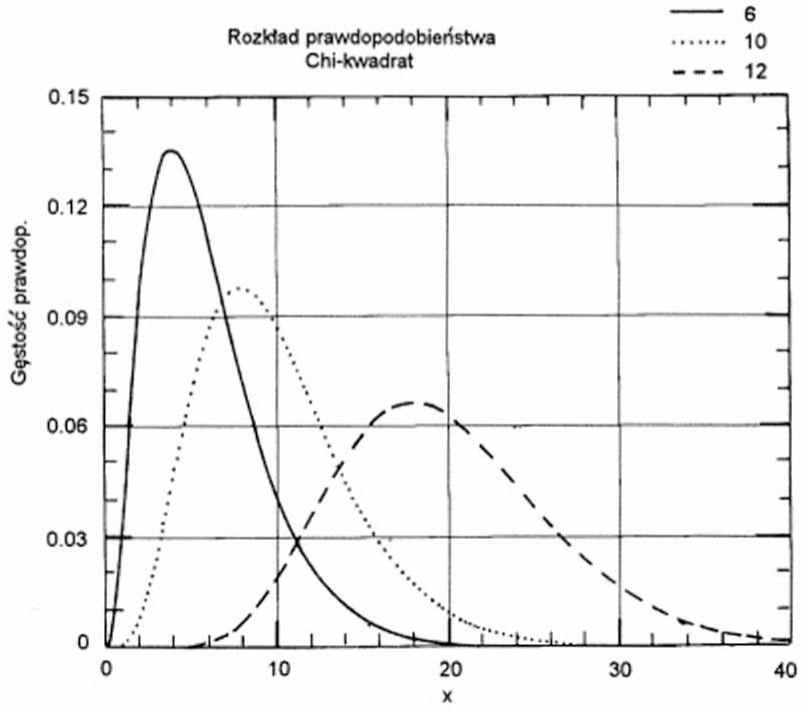
Rozkład χ^2 ma tylko jeden parametr — właśnie liczbę stopni swobody, dla której rezerwuje się specjalnie symbol ν.
Dowodzi się, że

$$E(\chi^2) = \nu \qquad \text{oraz} \qquad V(\chi^2) = 2\nu$$



Rys. D1.7 Rozkład χ^2.

Rozkład χ^2 jest przykładem rozkładu niesymetrycznego — dla $\nu = 1{,}2$ ma kształt typu J, dla $\nu \geq 3$ przybiera kształt skośnej krzywej dzwonowej, która wraz ze wzrostem ν staje się coraz bardziej symetryczna, dążąc do rozkładu normalnego (rysunki D1.7 oraz D1.8).



Rys. D1.8 Rozkład χ^2.

Można wykazać, że gdy zmienne losowe X_1, X_2, ..., X_n mają rozkłady $N(0, \sigma)$, wówczas statystyka[8]

$$\chi^2 = \frac{(n-1)\, s^2}{\sigma^2}$$

gdzie s^2 jest oszacowaniem wariancji, tzn.

$$s^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - E(X_i))^2 \ ,$$

8 Pod pojęciem statystyki rozumie się tu zmienną losową będącą funkcją innych zmiennych losowych

ma rozkład χ^2 o $\nu = n - 1$ stopniach swobody.

Rozkład χ^2 ma szerokie zastosowania szczególnie w testowaniu hipotez statystycznych jako miara ufności uzyskanego wyniku. Na rozkładzie χ^2 bazuje bardzo często stosowany test χ^2.

Test *t* Studenta

Odgrywający w statystyce matematycznej ogromną rolę rozkład podany w 1905 r. przez W.S. Gosseta o funkcji gęstości prawdopodobieństwa danej wzorem

$$f(t) = (\pi \nu)^{-1/2} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \nu^{-1} t^2\right)^{-\frac{1}{2}(\nu+1)}$$

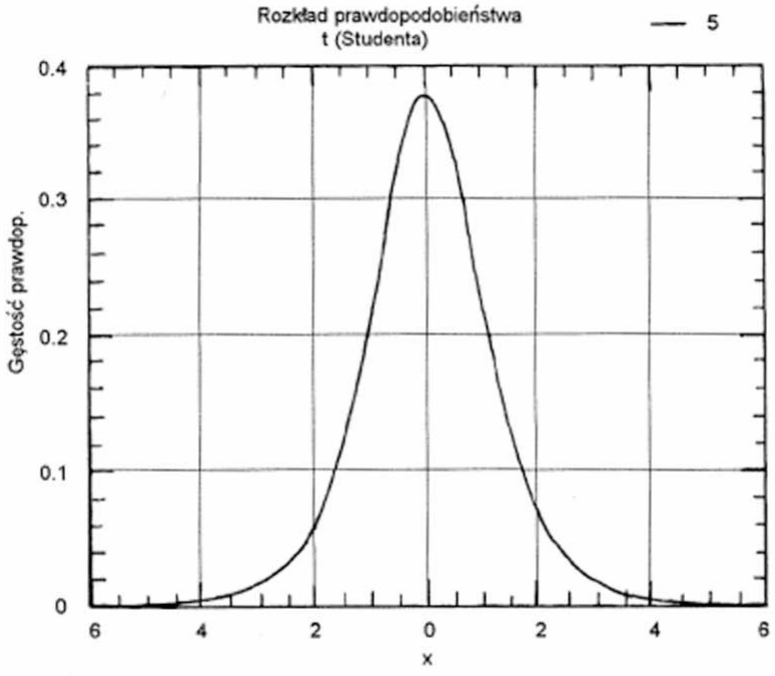
nosi nazwę rozkładu *t* Studenta[9]; jedynym parametrem tego rozkładu jest ν — liczba stopni swobody. Krzywą rozkładu *t* przedstawiono na rysunku D1.9. Rozkład *t* Studenta ma kształt zbliżony do rozkładu zmiennej normalnej, jednakże prawdopodobieństwa dużych odchyleń od osi symetrii ($t = 0$) są tu większe niż w rozkładzie normalnym. Rozkład zmiennej losowej *t* Studenta został stablicowany. Każdej ustalonej ilości stopni swobody $\nu = 1, 2, 3, \ldots$ odpowiada inny rozkład *t*. Stąd tablica prawdopodobieństw tego rozkładu obejmuje szereg wartości uzależnionych od ilości stopni swobody. Rozkład *t* wraz ze wzrostem ilości stopni swobody dąży do rozkładu normalnego $N(0, 1)$ — można się o tym przekonać porównując tablicę rozkładu *t* dla $n \geq 30$ z odpowiednią tablicą rozkładu normalnego.

Rozkład *t* to jeden z serii (obok rozkładów χ^2 i *F*) rozkładów z próby. Załóżmy bowiem, że niezależne zmienne losowe X_i ($i = 1, 2, \ldots, n$) mają jednakowy rozkład normalny $N(\mu, \sigma)$. Wartość średnia tych zmiennych $\bar{X}$ ma wtedy również rozkład normalny, ale o parametrach $N(\mu, \sigma/\sqrt{n})$. Rozkład ten można jednoznacznie określić tylko w przypadku, gdy znane jest odchylenie standardowe σ. Na ogół wartość σ nie jest znana i nie można przyjąć za nią wartości *s* obliczonej z próby

$$s = \sqrt{\frac{1}{n} \sum_{i=1} (X_i - \bar{X})^2}$$

---

9 Student to pseudonim naukowy Gosseta



Rys. D1.9 Rozkład *t*.

gdyż jest to jedynie pewne oszacowanie. Można jednakże wykazać, że statystyka zdefiniowana wzorem

$$t = \frac{\bar{X} - \mu}{s}\sqrt{n-1}$$

podlega rozkładowi *t* Studenta z $\nu = n - 1$ stopniami swobody.

Ze wzoru tego widać, że wartości statystyki *t* oblicza się na podstawie wyników doświadczalnych, gdy wartość μ jest z góry dana hipotetycznie.

Rozkład *F* Fishera-Snedecora

Funkcja gęstości rozkładu *F* ma bardzo skomplikowaną postać definicyjną — nie będziemy jej tu przytaczać, podamy natomiast twierdzenie które wiąże rozkład *F* ze statystyką z próby.

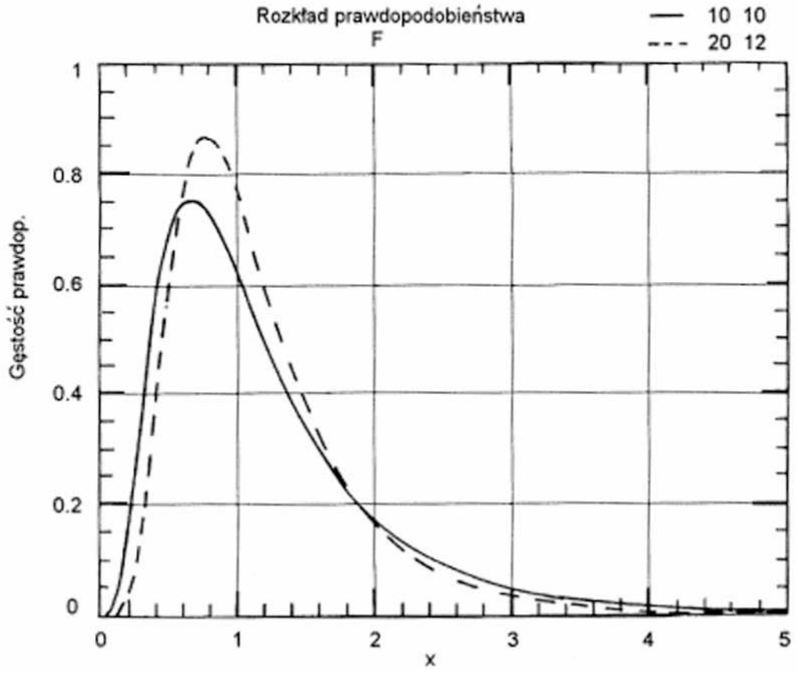
Rozpatrzmy mianowicie dwie grupy zmiennych losowych normalnych, z których pierwsza obejmuje ν_1 zmiennych $X_1, X_2, ..., X_{\nu_1}$ a druga ν_2 zmiennych $Y_1, Y_2, ..., Y_{\nu_2}$,

reprezentowanych przez wartości dwóch prób pobranych losowo z dwóch populacji. Dla każdej z tych grup[10] utwórzmy zmienną losową χ^2. Pierwsza z tych zmiennych χ_1^2 ma ν_1 stopni swobody, a druga ν_2 stopni swobody.
Wówczas funkcja postaci

$$F = \frac{\frac{1}{\nu_1}\sum_{i=1}^{\nu_1} X_i^2}{\frac{1}{\nu_2}\sum_{i=1}^{\nu_2} Y_i^2} = \frac{\frac{\chi_1^2}{\nu_1}}{\frac{\chi_2^2}{\nu_2}}$$

podlega rozkładowi F.

Rozkład F ma dwa parametry: ν_1, ν_2 nazywane liczbami stopni swobody tego rozkładu.



Rys. D1.10 Rozkład F.

10 zgodnie z definicją podaną w paragrafie omawiającym rozkład zmiennej losowej χ^2

Rozkład F przypomina swym kształtem rozkład χ^2; jest on różny od zera jedynie dla dodatnich wartości F; jest rozkładem asymetrycznym i szybko malejącym wraz ze wzrostem wartości F. Rysunek D1.10 przedstawia dwie krzywe rozkładu F dla różnych wartości stopni swobody.

Rozkład F został stablicowany, a tablica F przedstawia dla różnych par stopni swobody tylko dwie wartości, a mianowicie wartość 5-procentową i wartość 1-procentową[11]. Oznacza się je najczęściej symbolami $F_{0,05}$ oraz $F_{0,01}$. I tak odczytana z tablic dla $\nu_1 = 8$ oraz $\nu_2 = 14$ wartość $F_{0,05} = 2,70$ co oznacza, że prawdopodobieństwo tego iż wartość zmiennej F przekroczy 2,70 wynosi 0,05 = 5%.

Zmienna F nie zależy ani od średniej μ w populacji, ani od wariancji σ^2, a wartości tej zmiennej mogą być obliczane jedynie na podstawie wartości z próby.

Rozkłady empiryczne

Na koniec wspomnimy pokrótce o rozkładach empirycznych i ich związkach z rozkładami teoretycznymi.

Przypuśćmy, że dla określonej zmiennej losowej X, dyskretnej lub ciągłej, nie jest znany wzór przedstawiający rozkład prawdopodobieństwa — lub wzór jest znany, ale nie są znane występujące w nim stałe. Notując n losowo otrzymanych wartości X otrzymujemy próbę statystyczną. Dysponując próbą, możemy dla dowolnego przedziału (a, b) mieszczącego się w zbiorze wartości X podać częstość względną

$$\frac{n\,(a, b)}{n}$$

zdarzenia, że $a \leq X < b$. Chcąc przybliżyć rozkład X korzystamy z dokładnych wartości X i ich częstości, bądź też grupujemy obszar zmienności X na dowolną liczbę klas, wyliczając następnie częstości względne tych klas[12]. Otrzymamy w ten sposób empiryczny rozkład prawdopodobieństwa zwany także rozkładem w próbie. W przypadku danych pogrupowanych uważa się, że empiryczny rozkład prawdopodobieństwa daje dobry obraz rozkładu teoretycznego, gdy n jest tak duże, że liczba klas równych szerokości wynosi najmniej 15, a częstość każdej klasy wynosi najmniej 5.

Korzystając z częstości względnych możemy oprócz empirycznego rozkładu prawdopodobieństwa utworzyć również dystrybuantę empiryczną. Znane z teorii prawdopodobieństwa

---

11 celem uniknięcia pomyłek rozdziela się je często na dwie tablice

12 szczególnie wtedy, gdy próba jest bardzo liczna

twierdzenie Gliwenki określa, że gdy próba jest dostatecznie liczna, to możemy oczekiwać z prawdopodobieństwem bliskim jedności, że dystrybuanta empiryczna mało różni się od dystrybuanty teoretycznej.

Po utworzeniu empirycznego rozkładu prawdopodobieństwa możemy stosując odpowiednie testy statystyczne badać jego zgodność z określonym rozkładem teoretycznym. Innymi słowy, jesteśmy w stanie sprawdzić, czy odchyłki od rozkładu teoretycznego, jakie zaobserwowaliśmy w próbie mają jedynie charakter przypadkowy.

DODATEK 2.
PRZYKŁADOWE DANE

Poniżej zamieszczono przykładowe dane wykorzystywane jako podstawę obliczeń większości omawianych metod wielowymiarowych. Przykład ten został zaczerpnięty z książki H. Ahrensa i J. Lautera: Wielowymiarowa analiza wariancji.

Zamieszczone poniżej tabele dotyczą pacjentów pewnej kliniki cierpiących na nadczynność gruczołów tarczycowych. Na 23 osobach badanych wykonano odpowiednie pomiary, przy czym pacjentów podzielono na trzy klasy:

Klasa 1: Leczenie było pomyślne i po pewnym dłuższym czasie badania kliniczne wykazały, że pacjent jest wyleczony.

Klasa 2: Leczenie nie dało żadnych wyników pozytywnych, tzn. nadczynność gruczołów tarczycowych utrzymywała się nadal.

Klasa 3: Początkowo leczenie było pomyślne, ale przeprowadzone później badania kliniczne stwierdziły nawrót tej choroby.

Każdy pacjent został scharakteryzowany 10 cechami, oznaczonymi umownie $x1$, $x2$, ..., $x10$. Cechy $x1$, $x2$, $x3$, $x4$, $x5$ przedstawiają charakterystykę pacjenta przed rozpoczęciem terapii, a cechy $x6$, $x7$, $x8$, $x9$, $x10$ odpowiednie charakterystyki po zakończeniu terapii. Funkcjonowanie tarczycy badano przy pomocy izotopu jodu J-131 mierząc jego stężenie we krwi obwodowej po 3, 6, 24, i 48 godzinach po podaniu preparatu. W ten sposób otrzymano cechy $x1$, $x2$, $x3$, $x4$ oraz $x6$, $x7$, $x8$, $x9$. Cechy $x5$ oraz $x10$ oznaczają jod związany w białku proteinowym oznaczany po 48 godzinach po podaniu preparatu w okresie przed rozpoczęciem i po zakończeniu terapii.

Wartości liczbowe zebrano w dwóch poniższych tabelach.

Tab. D2.1

Numer pacjenta		Pierwszy test radiojodem				
		$x1$	$x2$	$x3$	$x4$	$x5$
Klasa 1	1	76,3	77,0	75,6	69,5	0,71
	2	56,7	54,6	49,2	40,5	2,20
	3	96,5	99,9	80,5	53,0	2,12
	4	54,4	94,1	89,5	84,0	1,40
	5	99,9	98,6	79,9	69,5	3,24
	6	99,9	99,9	98,5	69,5	2,79
	7	99,9	99,9	80,1	76,7	3,59
	8	73,6	72,0	66,7	50,2	1,90
	9	90,0	90,2	76,5	50,0	0,81
	10	95,2	77,6	76,4	74,0	2,54
	11	100,0	100,0	100,0	100,0	0,60
	12	100,0	100,0	100,0	100,0	1,31
	13	99,9	96,0	85,5	65,0	1,28
	14	100,0	92,9	94,0	78,0	1,93
	15	97,0	97,2	88,5	65,1	3,84
	16	89,6	100,0	100,0	86,8	0,08
Klasa 2	17	59,3	60,5	61,7	53,0	2,13
	18	83,4	86,2	79,3	78,0	1,04
	19	66,5	66,3	68,8	52,1	2,84
	20	96,4	100,4	98,0	94,0	1,33
Klasa 3	21	99,9	98,2	84,7	85,4	3,28
	22	77,0	81,0	63,6	54,2	3,34
	23	79,0	84,0	72,4	65,0	2,42

Tab. D2.2

Numer pacjenta		Drugi test radiojodem				
		$x6$	$x7$	$x8$	$x9$	$x10$
Klasa 1	1	14,4	14,3	25,8	25,1	0,20
	2	20,1	28,2	39,0	40,0	0,11
	3	24,1	29,1	40,6	32,1	0,17
	4	11,1	15,5	15,0	16,9	0,12
	5	16,3	24,4	34,0	32,1	0,36
	6	40,5	56,5	61,0	64,4	0,21
	7	52,7	58,2	48,6	50,0	0,53
	8	20,8	25,6	22,6	22,3	0,13
	9	14,0	10,3	8,0	3,1	0,18
	10	27,0	31,6	45,6	41,7	0,19
	11	44,3	59,2	66,3	63,8	0,22
	12	47,5	48,9	76,0	50,1	0,29
	13	54,0	58,0	61,7	57,0	0,19
	14	16,1	26,7	22,5	20,6	0,22
	15	57,5	61,2	76,0	74,5	0,49
	16	37,8	47,5	59,0	63,0	0,32
Klasa 2	17	55,8	58,5	51,5	48,0	2,74
	18	75,0	65,4	62,6	60,0	1,37
	19	72,0	81,3	75,5	65,0	0,70
	20	70,6	71,0	50,6	45,0	1,40
Klasa 3	21	24,1	33,3	45,5	45,0	0,22
	22	33,2	46,6	58,5	55,0	0,01
	23	30,4	35,6	46,2	44,6	0,09

DODATEK 3.
WIELOWYMIAROWY ROZKŁAD NORMALNY

Jak wiemy zmienna losowa x podlega rozkładowi normalnemu, co notujemy

$$x \equiv N(\mu,\sigma) ,$$

jeśli jej funkcja gęstości prawdopodobieństwa wyraża się wzorem

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \tag{1}$$

gdzie $E(x) = \mu$ jest średnią, natomiast $E(x-\mu)^2 = \sigma^2$ — wariancją zmiennej losowej x. Wzór (1) możemy zapisać w nieco innej, przydatnej do dalszych rozszerzeń postaci:

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{1}{2}(x-\mu)(\sigma^2)^{-1}(x-\mu)\right] \tag{2}$$

W przypadku dwóch zmiennych losowych x_1 i x_2 łączna gęstość rozkładu normalnego ma postać:

$$f(x) = \frac{1}{(2\pi)^{n/2}\,|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right] \tag{3}$$

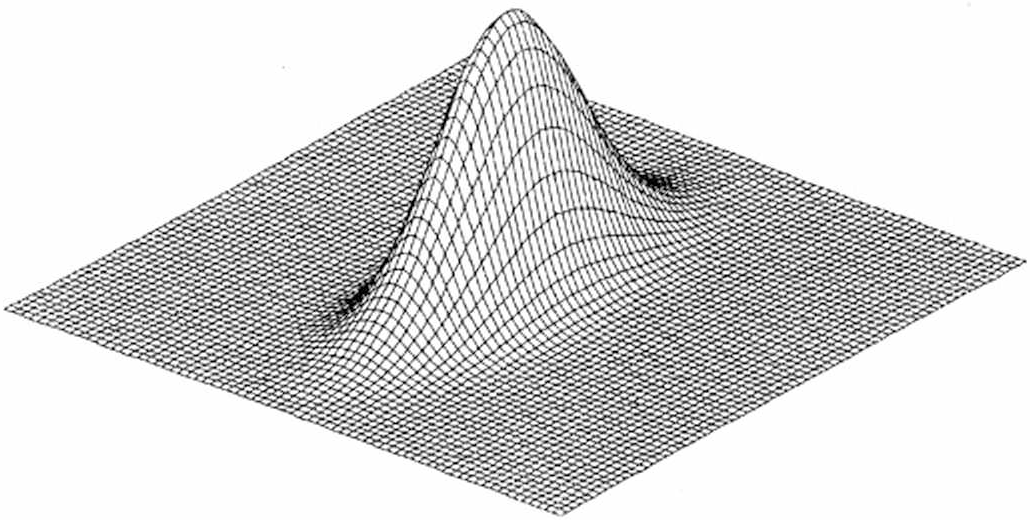
gdzie n jest równe liczbie zmiennych ($n = 2$), $x = [x_1, x_2]^T$ jest wektorem kolumnowym zmiennych losowych z wektorem średnich $\mu = [\mu_1, \mu_2]$, natomiast Σ jest macierzą kowariancji określoną dla dwóch zmiennych jako

$$\Sigma = \begin{bmatrix} E(x_1-\mu_1)^2 & E(x_1-\mu_1)(x_2-\mu_2) \\ E(x_1-\mu_1)(x_2-\mu_2) & E(x_1-\mu_1)^2 \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} \tag{4}$$

Występująca we wzorze (4) wartość $|\Sigma| = \sigma_1^2\,\sigma_2^2 - \sigma_{12}^2$ jest wyznacznikiem macierzy kowariancji.

Parametrami rozkładu są teraz: wektor średnich μ oraz macierz kowariancji Σ. Macierz ta zależy nie tylko od wariancji obu zmiennych, które znajdują się na głównej przekątnej, ale także od stopnia powiązań między nimi. Rozproszenie wyników w rozpatrywanej dwuwymiarowej przestrzeni zależy od stopnia powiązań osi układu współrzędnych, na których „rozpina się" przestrzeń naszych wektorów obserwacji. Jeśli korelacja między x_1 a x_2 jest duża, to najczęściej pojawiać się będą pary leżące wzdłuż linii regresji między x_1 a x_2. Zatem miejscami geometrycznymi punktów mających jednakową gęstość prawdopodobieństwa będą elipsy. Równania tych elips wyznacza się przyrównując wykładnik we wzorze (3) do wartości stałej.

Dwuwymiarowy rozkład normalny odpowiada powierzchni w przestrzeni trójwymiarowej (rysunek D3.1), której poziomymi przekrojami (poziomicami) są elipsy wzajemnie



Rys. D3.1.

koncentryczne. Dla maksymalnej wartości prawdopodobieństwa elipsa taka degeneruje się do punktu (μ_1, μ_2). Pionowe przekroje przechodzące przez środek rozkładu[1] mają postać jednowymiarowego rozkładu normalnego, którego szerokość jest wprost proporcjonalna do promienia elipsy kowariancji, branego wzdłuż linii danego przekroju. W szczególnym przypadku braku korelacji między x_1 a x_2, tzn. takim, dla którego $\text{cov}(x_1, x_2) = 0$, elipsy koncentracji przechodzą w okręgi o środku w punkcie (μ_1, μ_2).

Uogólnienie wzoru (1), lub raczej (2), na większą liczbę zmiennych (tzn. p zmiennych) polega na zastąpieniu jednej zmiennej losowej x wektorem losowym x o p składowych,

1 tzn. punkt o współrzędnych (μ_1, μ_2)

średniej μ tej zmiennej wektorem średnich μ i wreszcie wariancji σ^2 — macierzą kowariancji zmiennych $x_1, \ldots, x_p$.

Funkcja gęstości prawdopodobieństwa ma wówczas rozkład dany wzorem

$$f(x) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right] \tag{5}$$

Na podstawie tej funkcji gęstości stwierdzamy, że punkty x o jednakowej gęstości prawdopodobieństwa leżą na powierzchni elipsoidy w p-wymiarowej przestrzeni o równaniu

$$(x-\mu)^T \Sigma^{-1} (x-\mu) = c^2 \tag{6}$$

Elipsoida ta nazywa się elipsoidą koncentracji i stanowi uogólnienie granic $(\mu - c\sigma)$, $(\mu + c\sigma)$ obszaru rozproszenia jednowymiarowej zmiennej losowej o rozkładzie normalnym.

Identycznie jak poprzednio, rozkład normalny wielowymiarowej zmiennej losowej x notujemy jak poniżej:

$$x \cong N(\mu, \Sigma) \tag{7}$$

Rozkład normalny wielowymiarowy jest jednoznacznie określony, gdy dane są średnie wszystkich zmiennych oraz ich wariancje i kowariancje.

Przytoczymy jeszcze dwa ważne twierdzenia:

- TW 1. Jeżeli p-wymiarowa zmienna losowa x ma rozkład $N(\mu, \Sigma)$ i jeżeli U jest macierzą typu (p, u) rzędu u, to wówczas u-wymiarowa zmienna losowa

  $$z = U^T x$$

  ma rozkład normalny

  $$N(U^T \mu, U^T \Sigma U) \ .$$

- TW 2. Jeżeli p-wymiarowa zmienna losowa x ma rozkład $N(\mu, \Sigma)$ gdzie Σ jest macierzą kowariancji a $|\Sigma|$ oznacza wyznacznik tej macierzy, to prawdopodobieństwo, że między składowymi wektora x zachodzi przynajmniej jeden związek liniowy, jest równe jedności wtedy i tylko wtedy, gdy

  $$|\Sigma| = 0 \ .$$

Interpretacją tego twierdzenia jest, że jeśli $|\Sigma| = 0$, to cała masa prawdopodobieństwa leży w hiperpłaszczyźnie o liczbie wymiarów mniejszej niż p. Wyznacznik $|\Sigma|$ nosi nazwę wariancji uogólnionej. Nazwa ta wiąże się z pewnymi ważnymi własnościami macierzy Σ; macierz Σ jest mianowicie symetryczna i dodatnio określona, to znaczy forma kwadratowa

$$x\sum x^T = \sum_{i=1}^{p}\sum_{k=1}^{p}\sigma_{ik}x_i x_k$$

jest nieujemna dla wszystkich niezerowych wektorów x.

Stąd wynika między innymi, że jest stale

$$0 \leq |\sum| \leq \sigma_1^2 * \sigma_1^2 * \ldots * \sigma_p^2 .$$

Jeśli więc

$$|\sum| = |\sum|_{max} = \sigma_1^2 * \sigma_1^2 * \ldots * \sigma_p^2 ,$$

wówczas wszystkie kowariancje są zerami i zmienne losowe $x_1, x_2, \ldots, x_p$ są nieskorelowane. Jeżeli natomiast którekolwiek $\sigma_i = 0$, to wówczas rozkład x traci jeden wymiar.

DODATEK 4.
WYBRANE ZAGADNIENIA Z RACHUNKU MACIERZOWEGO

Metody wielowymiarowe którymi zajmowaliśmy się w tej części skryptu można przedstawić w sposób przejrzysty tylko za pomocą rachunku macierzowego. Zakładamy, że podstawy rachunku macierzowego są znane czytelnikowi, jednak pewne zależności i twierdzenia wykorzystywane w analizie wielowymiarowej zapewne nie były prezentowane w podstawowym kursie rachunku macierzowego. Dlatego też zebrane one zostały poniżej, przy czym z góry zastrzegamy się, że jest to bardzo wyrywkowy wybór.

D4. 1. Rząd macierzy A, co jest oznaczane jako rz A, jest to maksymalna liczba liniowo niezależnych wierszy (lub kolumn) macierzy A traktowanych jako wektory. Oczywiście

$$\text{rz } A_{(m,n)} \leq \min(m, n) \ .$$

D4. 2. Śladem macierzy kwadratowej A_n nazywa się sumę elementów diagonalnych tej macierzy, tzn.:

$$\text{tr}(A) = \sum_{i=1}^{n} a_{ii}$$

D4. 3. Macierz kwadratowa nazywa się macierzą diagonalną, jeżeli wszystkie jej elementy niediagonalne są równe zeru.

D4. 4. Macierz kwadratową A nazywa się macierzą ortogonalną jeżeli:

$$A^T A = A A^T = I \ .$$

D4. 5. Dla każdej macierzy ortogonalnej A zachodzi

$$A^T = A^{-1}$$

oraz

$|A| = 1$ lub $|A| = -1$.

D4. 6. Macierz kwadratową A nazywamy symetryczną, gdy

$$A = A^T .$$

D4. 7. Dla macierzy symetrycznej A stopnia n można znaleźć taką macierz ortogonalną C, że

$$C^T A C = L$$

gdzie

$$L = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \lambda_n \end{bmatrix}$$

jest macierzą diagonalną o elementach rzeczywistych ustawionych w porządku malejącym

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n .$$

D4. 8. Dla danej macierzy symetrycznej A stopnia n macierz diagonalna L, tzn. liczby $\lambda_1, \lambda_2, \dots, \lambda_n$, są też określone jako rozwiązania równania charakterystycznego macierzy

$$|A - \lambda J| = 0 .$$

Liczby $\lambda_1, \lambda_2, \dots, \lambda_n$ noszą nazwę wartości własnych macierzy A. Dla każdej wartości własnej λ_i macierzy A istnieje tzw. wektor własny (lub charakterystyczny) x_i, taki że:

$$A x_i = \lambda_i x_i , \qquad \text{przy czym} \qquad x_i^T x_i = 1$$

Gdy $\lambda \neq \lambda$ wtedy wektory własne są wzajemnie ortogonalne (prostopadłe), tzn.:

$$x_i^T x_j = 0 \qquad \text{lub} \qquad x_j^T x_i = 0 .$$

Problem rozwiązania równania $Ax = \lambda x$ nazywa się zagadnieniem wartości własnych.

D4. 9. Liczba niezerowych wartości własnych macierzy A pokrywa się z rzędem macierzy A.

D4.10. Dla każdej macierzy symetrycznej A zachodzą równości:

$$\text{tr}\,(A) = \sum_{i=1}^{n} \lambda_i$$

$$|A| = \prod_{i=1}^{n} \lambda_i$$

D4.11. Formą kwadratową o macierzy A nazywamy wyrażenie postaci

$$x^T Ax = \sum_{i,j=1}^{n} a_{ij} x_i x_j, \qquad \text{gdzie} \qquad x \neq 0$$

Jeżeli $x^TAx > 0$, to mówimy że macierz A jest dodatnio określona i wszystkie wartości własne macierzy A są dodatnie; jeżeli $x^TAx \geq 0$, mówimy że macierz A jest dodatnio półokreślona i wszystkie wartości własne macierzy A są nieujemne. Formy kwadratowe, których argumentami są zmienne losowe X_i odgrywają ważną rolę zarówno w jednowymiarowej, jak i w wielowymiarowej statystyce. Na przykład, suma kwadratów odchyleń od średniej próbkowej

$$\sum_{i=1}^{N} (x_i - \bar{x})^2 = \sum_{i=1}^{N} x_i^2 - \frac{1}{N} (\sum x_i)^2$$

może być zapisana w postaci formy kwadratowej obserwacji x_i o macierzy

$$A = \begin{bmatrix} \frac{N-1}{N} & -\frac{1}{N} & \cdots & -\frac{1}{N} \\ -\frac{1}{N} & \frac{N-1}{N} & \cdots & -\frac{1}{N} \\ \cdots & \cdots & \cdots & \cdots \\ -\frac{1}{N} & -\frac{1}{N} & \cdots & \frac{N-1}{N} \end{bmatrix}$$

Z form kwadratowych korzystamy w analizie wariancji i w tzw. obszarach ufności. Zmienna losowa T^2 Hotellinga jest taką formą; elipsoida ufności również wyraża się przy użyciu formy kwadratowej.

D4.12. Dla każdych dwóch macierzy symetrycznych A i B typu (n, n), gdzie macierz B jest dodatnio określona istnieje macierz C taka, że:

$$C^T A\ C = L \qquad \text{oraz} \qquad C^T B\ C = I\ ,$$

gdzie L jest macierzą diagonalną

$$L = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \lambda_n \end{bmatrix}$$

czyli dla danych macierzy A i B macierz L jest jednoznacznie wyznaczona. Liczby $\lambda_1, \lambda_2, \dots, \lambda_n$ można otrzymać także jako rozwiązania równania charakterystycznego

$$|A - \lambda B| = 0 \qquad \text{lub} \qquad |AB^{-1} - I| = 0\ ;$$

są to wartości własne macierzy AB^{-1}.
Dla każdej wartości własnej λ_i macierzy AB^{-1} istnieje wektor własny x_i taki, że

$$A\ x_i = \lambda_i B\ x_i \qquad \text{oraz} \qquad x_i^T B\ x_i = 1\ .$$

Gdy $\lambda_i \neq \lambda$, wtedy $x_i^T B\ x_i = 0$.
Dla macierzy A i B zachodzą równości:

$$\text{tr}\,(AB^{-1}) = \sum_{i=1}^{n} \lambda_i\ ,$$

$$|AB^{-1}| = \prod_{i=1}^{n} \lambda_i\ ,$$

Jeżeli macierz A jest dodatnio określona[1], to wartości własne równania

$$Ax = \lambda Bx$$

---

1 macierz B jest z założenia dodatnio określona

są dodatnie; natomiast jeżeli macierz A jest dodatnio półokreślona, to wartości własne tego równania są nieujemne.

D4.13. Macierz symetryczna A nazywa się idempotentną gdy

$$A\,A = A^2 = A$$

Macierz idempotentna jest dodatnio półokreślona. Jest ona dodatnio określona wtedy i tylko wtedy, gdy jest macierzą jednostkową. Macierz idempotentna ma wartości własne równe 1 lub 0. Stąd też mamy równość:

$$\mathrm{tr}(A) = \mathrm{rz}\ A$$

Forma kwadratowa, której macierz jest idempotentna, może być zredukowana do sumy kwadratów n zmiennych. Łatwo można sprawdzić, że macierz A z punktu D4.11 jest idempotentna oraz że można dokonać następującego przekształcenia:

$$\sum_{i=1}^{n} (x_i - \overline{x})^2 = y_1^2 + \ldots + y_{N-1}^2$$

Statystyczna niezależność nowych zmiennych losowych wynika z ortogonalności przekształcenia.

D4.14. Często wygodnie jest zapisywać macierze w postaci blokowej przez zestawienie dwóch lub większej liczby pewnych innych macierzy. Macierze blokowe mają na przykład postać:

$$A = \left[\begin{array}{cc} P & Q \\ \hdashline R & S \end{array}\right] = \begin{bmatrix} P & Q \\ R & S \end{bmatrix}$$

przy czym linie przerywane opuszcza się, gdy nie ma obawy o nieporozumienia. W powyższym wzorze macierze P i Q mają jednakową liczbę wierszy, macierze P i R mają jednakową liczbę kolumn, itd. Z definicji macierzy transponowanej wynika, że:

$$A^T = \left[\begin{array}{c:c} P^T & R^T \\ \hdashline Q^T & S^T \end{array}\right]$$

Iloczyn dwóch macierzy blokowych można otrzymać za pomocą reguł mnożenia traktując macierze jako elementy, np.:

$$AC = \begin{bmatrix} P & Q \\ R & S \end{bmatrix} \begin{bmatrix} E & F \\ G & H \end{bmatrix} = \begin{bmatrix} PE + QG & PF + QH \\ RE + SG & RF + SH \end{bmatrix}$$

zakładając, że iloczyny PE i inne istnieją.

D4.15. Jeśli nieosobliwa macierz kwadratowa A ma postać

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

gdzie podmacierze A_{11} i A_{22} są kwadratowe, a ponadto macierze A_{22} i

$$A_{11.2} = A_{11} - A_{12} A_{22}^{-1} A_{21}$$

są nieosobliwe, to na macierz odwrotną A^{-1} mamy następujący wzór

$$A^{-1} = \begin{bmatrix} A_{11.2}^{-1} & -A_{11.2}^{-1} A_{12} A_{22}^{-1} \\ -A_{22}^{-1} A_{21} A_{11.2}^{-1} & A_{22}^{-1} + A_{22}^{-1} A_{21} A_{11.2}^{-1} A_{12} A_{22}^{-1} \end{bmatrix}$$

W zasadzie więc dla obliczenia macierzy odwrotnej do A, tzn. A^{-1}, wystarczy odwrócić jej podmacierz A_{22} oraz pewną funkcję jej podmacierzy, tzn. $A_{11.2}$.

LITERATURA

W dodatku zestawiono pozycje łatwiej osiągalne i za jednym wyjątkiem jedynie książkowe, oferujące bardziej ogólne prezentacje problemów biometrii ze szczególnym uwzględnieniem analizy wielowymiarowej.

1. Ahrens H., Läuter J.: Wielowymiarowa analiza wariancji, PWN, Warszawa, 1979
2. Anderberg M. R.: Cluster Analysis for Applications, Academic Press, New York, 1973
3. Anderson W. T.: An Introduction to Multivariate Statistical Analysis, Wiley, New York, 1958
4. Armitage P.: Metody statystyczne w badaniach medycznych, PZWL, Warszawa, 1978
5. Blalock H. M.: Statystyka dla socjologów, PWN, Warszawa, 1977
6. Burzyński J.: Kurs rachunku prawdopodobieństwa i statystyki matematycznej, AGH, Kraków, 1976
7. Classification Pattern Recognition and Reduction of Dimensionality, ed. P. R. Krisnaiah, L. N. Kanal, North-Holland, Amsterdam, 1982
8. Domański C.: Statystyczne testy nieparametryczne, PWE, Warszawa, 1979
9. Domański C.: Testy statystyczne, PWE, Warszawa, 1990
10. Draper R. N., Smith H.: Analiza regresji stosowana, PWN, Warszawa, 1973
11. Freund J. E.: Podstawy nowoczesnej statystyki, PWE, Warszawa, 1968
12. Gnanadesikan R.: Methods of Statistical Data Analysis of Multivariate Observations, New York, John Wiley, 1977
13. Greń J.: Statystyka matematyczna. Modele i zadania, PWN, Warszawa, 1982
14. Harris R. J.: A Primer of Multivariate Statistics, New York, Academic Press, 1975
15. Hartigan J. A.: Clustering Algorithms, Wiley, New York, 1975
16. Krzyśko M., Ratajczak W.: Analiza kanoniczna, Listy Biometryczne, 65-67, 1-46, 1978
17. Marek T., Noworol C.: Wprowadzenie do wielozmiennowej analizy regresji, skrypt UJ, Kraków, 1985
18. Morrison D. F.: Multivariate statistical methods, New York, McGraw — Hill, 1967

19. Oktaba W.: Elementy statystyki matematycznej i metodyka doświadczalnictwa, PWN, Warszawa, 1966
20. Oktaba W.: Metody statystyki matematycznej w doświadczalnictwie, PWN, Warszawa, 1980
21. Parker R. E.: Wprowadzenie do statystyki dla biologów, PWN, Warszawa, 1978
22. Rao R. C.: Modele liniowe statystyki matematycznej, PWN, Warszawa, 1982
23. Späth H.: Cluster Analysis Algorithms for Data Reduction and Classification of Object, Ellis Horwood, Chichester, 1982
24. Steczkowski J., Woźniak M., Zając K., Zeliaś A.: Statystyka matematyczna w zastosowaniach, AE, Kraków, 1979
25. Tatsuoka M. M.: Multivariate Analysis, Wiley, New York, 1971
26. Wielozmiennowe modele statystyczne w badaniach psychologicznych, pod red. J. Brzezińskiego, PWN, Warszawa, 1987
27. Zieliński R.: Tablice statystyczne, PWN, Warszawa, 1974
28. Żuk B.: Biometria stosowana, PWN, Warszawa, 1989

WYBRANE TABLICE STATYSTYCZNE

W dodatku zebrano najczęściej wykorzystywane w praktyce tablice statystyczne ograniczając się do tych, o których jest mowa w skrypcie.

Prezentowany wybór zawiera tablice:

1. Prawdopodobieństwa $Q(n; k, p)$ w rozkładzie dwumianowym (Bernoulliego), określone jako

$$Q(n; k, p) = \sum_{i=k}^{n} P\{X = i\}$$

gdzie

$$P\{X = k\} = \binom{n}{k} p^k (1-p)^{n-k}$$

2. Wartości funkcji prawdopodobieństwa rozkładu Poissona

$$p_k = P\{X = k\} = \frac{\lambda^k}{k!} e^{-\lambda}$$

3. Prawdopodobieństwa $Q(k; \lambda)$ rozkładu Poissona określone jako

$$Q(k; \lambda) = \sum_{i=k}^{\infty} P\{X = i\}$$

gdzie

$$P\{X = k\} = \frac{\lambda^k}{k!} e^{-\lambda}$$

4. Dystrybuanta standardowego rozkładu normalnego $N(0, 1)$ oznaczona jako $\Phi(x)$ dla $x \geq 0$.

Wartości dla $x \leq 0$ wyliczamy z zależności

$$\Phi(x) = 1 - \Phi(-x)$$

5. Wartości kwantyli rozkładu normalnego $N(0, 1)$ rzędu $\alpha \geq 0{,}5$, tzn. takie liczby x_α, że $\Phi(x_\alpha) = \alpha$.
   Dla $\alpha < 0{,}5$ wartości kwantyli wyliczamy ze wzoru

$$x_\alpha = -x_{1-\alpha}$$

6. Wartości krytyczne χ^2_α testu jednostronnego opartego na rozkładzie χ^2, spełniajce warunek

$$P\left\{\chi^2 \geq \chi^2_\alpha\right\} = \alpha.$$

   Dla testu dwustronnego na poziomie istotności α za górną i dolną wartość krytyczną przyjmuje się wartości odpowiadajce prawdopodobieństwom $\alpha/2$ i $(1 - \alpha/2)$.

7. Wartości krytyczne t_α testu dwustronnego opartego na rozkładzie t Studenta, spełniające warunek

$$P\left\{|t| > t_\alpha\right\} = \alpha$$

   Dla testu jednostronnego t_α odpowiada poziomowi istotności $\alpha/2$.

8. Wartości krytyczne $F_{\alpha, n, m}$ testu jednostronnego opartego na rozkładzie F Snedecora, spełniające warunek

$$P\left\{F > F_{\alpha, n, m}\right\} = \alpha$$

   dla wartości $\alpha = 0{,}05$ oraz $\alpha = 0{,}01$.

9. Wartości krytyczne wsółpczynnika korelacji r dla weryfikacji hipotezy $H_0 : \rho = 0$ przy hipotezie alternatywnej $H_1 : \rho \neq 0$ (test dwustronny).

10. Wartości krytyczne współczynnika korelacji wielokrotnej $R(\alpha; k, \nu)$ weryfikacji hipotezy $H_0 : \rho = 0$ przy hipotezie alternatywnej $H_1 : \rho \neq 0$, gdzie k jest liczbą zmiennych (łącznie ze zmienną zależną), $\nu = n - k$, natomiast n jest licznością próby.

11. Wartości krytyczne u_α serii Walda-Wolfowitza

$$P\left\{U \leq u_\alpha\right\} \leq \alpha = 0{,}05$$

12. Wartości krytyczne u_α testu rang Wilcoxona-Manna-Whitneya

$$P\{U \leq u_\alpha\} \leq \alpha = 0{,}05$$

13. Wartości krytyczne $D_n(\alpha)$ w teście Kołmogorowa.
14. Wartości krytyczne $mnD_{m,n}(\alpha)$ w teście Smirnowa ($n \neq m$) zgodności dwóch rozkładów.
15. Wartości krytyczne $nD_{n,n}(\alpha)$ w teście Smirnowa ($n \neq m$) zgodności dwóch rozkładów.
16. Wartości wielomianów ortogonalnych $\xi_p(x)$, gdzie p jest stopniem wielomianu.
17. Stablicowane wartości przekształcenia

$$z = \frac{1}{2} \ln \frac{1+r}{1-r}$$

18. Stablicowane wartości przekształcenia

$$r = \frac{e^{2z} - 1}{e^{2z} + 1}$$

Prawdopodobieństwo $Q(k; n, p)$ w rozkładzie dwumianowym

n	k	p 0,01	0,05	0,10	0,20	0,30	0,40	0,50
2	0	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000
	1	0,01990	0,09750	0,19000	0,36000	0,51000	0,64000	0,75000
	2	,00010	,00250	,01000	,04000	,09000	,16000	,25000
3	0	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000
	1	0,02970	0,14263	0,27100	0,48800	0,65700	0,78400	0,87500
	2	,00030	,00725	,02800	,10400	,21600	,35200	,50000
	3	,00000	,00013	,00100	,00800	,02700	,06400	,12500
4	0	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000
	1	0,03940	0,18549	0,34390	0,59040	0,75990	0,87040	0,93750
	2	,00059	,01402	,05230	,18080	,34830	,52480	,68750
	3	,00000	,00048	,00370	,02720	,08370	,17920	,31250
	4		,00001	,00010	,00160	,00810	,02560	,06250
5	0	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000
	1	0,04901	0,22622	0,40951	0,67232	0,83193	0,92224	0,96875
	2	,00098	,02259	,08146	,26272	,47178	,66304	,81250
	3	,00001	,00116	,00856	,05792	,16308	,31744	,50000
	4	,00000	,00003	,00046	,00672	,03078	,08704	,18750
	5		,00000	,00001	,00032	,00243	,01024	,03125
6	0	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000
	1	0,05852	0,26491	0,46856	0,73786	0,88235	0,95334	0,98438
	2	,00146	,03277	,11427	,34464	,57983	,76672	,89063
	3	,00002	,00223	,01585	,09888	,25569	,45568	,65625
	4	,00000	,00009	,00127	,01696	,07047	,17920	,34375
	5		,00000	,00006	,00160	,01094	,04096	,10938
	6			,00000	,00006	,00073	,00410	,01563
7	0	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000
	1	0,06793	0,30166	0,52170	0,79028	0,91765	0,97201	0,99219
	2	,00203	,04438	,14969	,42382	,67058	,84137	,93750
	3	,00003	,00376	,02569	,14803	,35293	,58010	,77344
	4	,00000	,00019	,00273	,03334	,12604	,28979	,50000
	5		,00001	,00018	,00467	,02880	0,9626	,22656
	6		,00000	,00001	,00037	,00379	,01884	,06250
	7			,00000	,00001	,00022	,00164	,00781
8	0	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000
	1	0,07726	0,33658	0,56935	0,83223	0,94235	0,98320	0,99609
	2	,00269	,05724	,18690	,49668	,74470	,89362	,96484
	3	,00005	,00579	,03809	,20308	,44823	,68461	,85547
	4	,00000	,00037	,00502	,05628	,19410	,40591	,63672
	5		,00002	,00043	,01041	,05797	,17367	,36328
	6		,00000	,00002	,00123	,01129	,04981	,14453
	7			,00000	,00008	,00129	,00852	,03516
	8				,00000	,00007	,00066	,00391

n	k	p 0,01	0,05	0,10	0,20	0,30	0,40	0,50
9	0	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000
	1	0,08648	0,36975	0,61258	0,86578	0,95965	0,98992	0,99805
	2	,00344	,07121	,22516	,56379	,80400	,92946	,98047
	3	,00008	,00836	,05297	,26180	,53717	,76821	,91016
	4	,00000	,00064	,00833	,08564	,27034	,51739	,74609
	5		,00003	,00089	,01958	,09881	,26657	,50000
	6		,00000	,00006	,00307	,02529	,09935	,25391
	7			,00000	,00031	,00429	,02503	,08984
	8				,00002	,00043	,00380	,01953
	9				,00000	,00002	,00026	,00195
10	0	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000
	1	0,09562	0,40126	0,65132	0,89263	0,97175	0,99395	0,99902
	2	,00427	,08614	,26390	,62419	,85069	,95364	,98926
	3	,00011	,01150	,07019	,32220	,61722	,83271	,94531
	4	,00000	,00103	,01280	,12087	,35039	,61772	,82813
	5		,00006	,00163	,03279	,15027	,36690	,62305
	6		,00000	,00015	,00637	,04735	,16624	,37695
	7			,00001	,00086	,01059	,05476	,17188
	8			,00000	,00008	,00159	,01229	,05469
	9				,00000	,00014	,00168	,01074
	10					,00001	,00010	,00098
12	0	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000
	1	0,11362	0,45964	0,71757	0,93128	0,98616	0,99782	0,99976
	2	,00617	,11836	,34100	,72512	,91497	,98-41	,99683
	3	,00021	,01957	,11087	,44165	,74718	,91656	,98071
	4	,00000	,00224	,02564	,20543	,50748	,77466	,92700
	5		,00018	,00433	,07256	,27634	,56182	,80615
	6		,00001	,00054	,01941	,11785	,33479	,61279
	7		,00000	,00005	,00390	,03860	,15821	,38721
	8			,00000	,00058	,00949	,05731	,19385
	9				,00006	,00169	,01527	,07300
	10				,00000	,00021	,00281	,01929
	11					,00002	,00032	,00317
	12					,00000	,00002	,00024

n	k	p 0,01	0,05	0,10	0,20	0,30	0,40	0,50
15	0	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000
	1	0,13994	0,53671	0,79411	0,96482	0,99525	0,99953	0,99997
	2	,00963	,17093	,45096	,83287	,96473	,99483	,99951
	3	,00042	,03620	,18406	,60198	,87317	,97289	,99631
	4	,00001	,00547	,05556	,35184	,70313	,90950	,98242
	5	,00000	,00061	,01272	,16423	,48451	,78272	,94077
	6		,00005	,00225	,01806	,27838	,59678	,84912
	7		,00000	,00031	,00424	,13114	,39019	,69638
	8			,00003	,00078	,05001	,21310	,50000
	9			,00000	,00011	,01524	,09505	,30362
	10				,00001	,00365	,03383	,15088
	11				,00000	,00067	,00935	,05923
	12					,00009	,00193	,01758
	13					,00001	,00028	,00369
	14					,00000	,00003	,00049
	15						,00000	,00003
20	0	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000	
	1	0,18209	0,64151	0,87842	0,98847	0,99920	0,99996	1,00000
	2	,01686	,26416	,60825	,93082	,99236	,99948	0,99998
	3	,00100	,07548	,32307	,79392	,96452	,99639	,99980
	4	,00004	,01590	,13295	,58855	,89291	,98404	,99871
	5	,00000	,00257	,04317	,37035	,76249	,94905	,99409
	6		,00033	,01125	,19579	,58363	,87440	,97931
	7		,00003	,00239	,08669	,39199	,74999	,94234
	8		,00000	,00042	,03214	,22773	,58411	,86841
	9			,00006	,00998	,11333	,40440	,74828
	10			,00001	,00259	,04796	,24466	,58810
	11			,00000	,00056	,01714	,12752	,41190
	12				,00010	,00514	,05653	,25172
	13				,00002	,00128	,02103	,13159
	14				,00000	,00026	,00647	,05756
	15					,00004	,00161	,02069
	16					,00001	,00032	,00591
	17					,00000	,00005	,00129
	18						,00001	,00020
	19						,00000	,00002
	20							,00000

n	k	p 0,01	0,05	0,10	0,20	0,30	0,40	0,50
30	0	1,00000	1,00000	1,00000	1,00000	1,00000		
	1	0,26030	0,98536	0,95761	0,99876	0,99998		
	2	,03615	,44646	,81630	,98948	,99969	1,00000	
	3	,00332	,18782	,58865	,95582	,99789	0,99995	
	4	,00022	,06077	,35256	,87729	,99068	,99969	1,00000
	5	,00001	,01564	,17549	,74477	,96985	,99849	0,99997
	6	,00000	,00328	,07319	,57249	,92341	,99434	,99984
	7		,00057	,02583	,39303	,84048	,98282	,99928
	8		,00008	,00778	,23921	,71862	,95648	,99739
	9		,00001	,00202	,12865	,56848	,90599	,99194
	10		,00000	,00045	,06109	,41119	,82371	,97861
	11			,00009	,02562	,26963	,70853	,95063
	12			,00002	,00949	,15932	,56891	,89976
	13			,00000	,00311	,08447	,42153	,81920
	14				,00090	,04005	,28550	,70767
	15				,00023	0,1694	,17537	,57223
	16				,00005	,00637	,09706	,42777
	17				,00001	,00212	,04811	,29233
	18				,00000	,00063	,02124	,18080
	19					,00016	,00830	,10024
	20					,00004	,00285	,04937
	21					,00001	,00086	,02139
	22					,00000	,00022	,00806
	23						,00005	,00261
	24						,00001	,00072
	25						,00000	,00016
	26							,00003
	27							,00000

Tablica 2

Rozkład Poissona — wartości funkcji prawdopodobieństwa $p_k = \frac{\lambda^k}{k!} e^{-\lambda}$

k \ λ	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1,0
0	0,9048	,8187	,7408	,6703	,6065	,5488	,4966	,4493	,4066	,3679
1	,0905	,1637	,2222	,2681	,3033	,3293	,3476	,3595	,3659	,3670
2	,0045	,0164	,0333	,0536	,0758	,0988	,1217	,1438	,1647	,1839
3	,0002	,0011	,0033	,0072	,0126	,0198	,0284	,0383	,0494	,0613
4		,0001	,0003	,0007	,0016	,0030	,0050	,0077	,0111	,0153
5				,0001	,0002	,0004	,0007	,0012	,0020	,0031
6							,0001	,0002	,0003	,0005
7										,0001

k \ λ	1,1	1,2	1,3	1,4	1,5	1,6	1,7	1,8	1,9	2,0
0	0,3329	,3012	,2725	,2466	,2231	,2019	,1827	,1653	,1496	,1353
1	,3662	,3614	,3543	,3452	,3347	,3230	,3106	,2975	,2842	,2707
2	,2014	,2169	,2303	,2417	,2510	,2584	,2640	,2678	,2700	,2707
3	,0738	,0867	,0998	,1128	,1255	,1378	,1496	,1607	,1710	,1804
4	,0203	,0260	,0324	,0395	,0471	,0551	,0636	,0723	,0812	,0902
5	,0045	,0062	,0084	,0111	,0141	,0176	,0216	,0260	,0309	,0361
6	,0008	,0012	,0018	,0026	,0035	,0047	,0061	,0078	,0098	,0120
7	,0001	,0002	,0003	,0005	,0008	,0011	,0015	,0020	,0027	,0034
8			,0001	,0001	,0001	,0002	,0003	,0005	,0006	,0009
9							,0001	,0001	,0001	,0002

k \ λ	2,1	2,2	2,3	2,4	2,5	2,6	2,7	2,8	2,9	3,0
0	0,1225	,1108	,1003	,0907	0,821	0,743	,0672	,0608	,0550	0,498
1	,2572	,2438	,2306	,2177	,2052	,1931	,1815	,1703	,1596	,1494
2	,2700	,2681	,2652	,2613	,2565	,2510	,2450	,2384	,2314	,2240
3	,1890	,1966	,2033	,2090	,2138	,2176	,2205	,2225	,2237	,2240
4	,0992	,1082	,1169	,1254	,1336	,1414	,1488	,1557	,1622	,1680
5	,0417	,0476	,0538	,0602	,0668	,0735	,0804	,0872	,0940	,1008
6	,0146	,0174	,0206	,0241	,0278	,0319	,0362	,0407	,0455	,0504
7	,0044	,0055	,0068	,0083	,0099	,0118	,0139	,0163	,0188	,0216
8	,0011	,0015	,0019	,0025	,0031	,0038	,0047	,0057	,0068	,0081
9	,0003	,0004	,0005	,0007	,0009	,0011	,0014	,0018	,0022	,0027
10	,0001	,0001	,0001	,0002	,0002	,0003	,0004	,0005	,0006	,0008
11						,0001	,0001	,0001	,0002	,0002
12										,0001

k \ λ	3,5	4,0	4,5	5,0	6,0	7,0	8,0	9,0	10,0
0	0,0302	,0183	,0111	,0067	,0025	,0009	,0003	,0001	,0000
1	,1057	,0733	,0500	,0337	,0149	,0064	,0027	,0011	,0005
2	,1850	,1465	,1125	,0842	,0449	,0223	,0107	,0050	,0023
3	,2158	,1954	,1687	,1404	,0892	,0521	,0286	,0150	,0076
4	,1888	,1954	,1898	,1755	,1339	,0912	,0573	,0337	,0189
5	,1322	,1563	,1708	,1755	,1606	,1277	,0916	,0607	,0378
6	,0771	,1042	,1281	,1462	,1606	,1490	,1221	,0911	,0631
7	,0385	,0595	,0823	,1044	,1377	,1490	,1396	,1171	,0901
8	,0169	,0298	,0463	,0653	,1033	,1304	,1369	,1318	,1126
9	,0066	,0132	,0232	,0363	,0688	,1014	,1241	,1318	,1251
10	,0023	,0053	,0104	,0181	,0413	,0710	,0993	,1186	,1251
11	,0007	,0019	,0043	,0082	,0225	,0452	,0722	,0970	,1137
12	,0002	,0006	,0016	,0034	,0113	,0264	,0481	,0721	,0948
13	,0001	,0002	,0006	,0013	,0052	,0142	,0296	,0504	,0729
14		,0001	,0002	,0005	,0022	,0071	,0169	,0324	,0521
15			,0001	,0002	,0009	,0033	,0090	,0194	,0347
16					,0003	,0014	,0045	,0109	,0217
17					,0001	,0006	,0021	,0058	,0128
18						,0002	,0009	,0029	,0071
19						,0001	,0004	,0014	,0037
20							,0002	,0006	,0019
21							,0001	,0003	,0009
22								,0001	,0004
23									,0002

Tablica 3

Prawdopodobieństwa $Q(k; \lambda)$ w rozkładzie Poissona

k \ λ	0,5	0,6	0,7	0,8	0,9	1,0	1,2	1,4	1,6	1,8
0	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
1	0,3935	0,4512	0,5034	0,5507	0,5934	0,6321	0,6988	0,7534	0,7981	0,8347
2	,0902	,1219	,1558	,1912	,2275	,2642	,3374	,4082	,4751	,5372
3	,0144	,0231	,0341	,0474	,0629	,0803	,1205	,1665	,2166	,2694
4	,0018	,0034	,0058	,0091	,0135	,0190	,0338	,0537	,0788	,1087
5	,0002	,0004	,0009	,0014	,0023	,0037	,0077	,0143	,0237	,0364
6	,0000	,0000	,0001	,0002	,0003	,0006	,0015	,0032	,0060	0,104
7			,0000	,0000	,0000	,0001	,0003	,0006	,0013	,0026
8						,0000	,0000	,0001	,0003	,0006
9								,0000	,0000	,0001
10										,0000

k \ λ	2,0	2,5	3,0	3,5	4,0	4,5	5,0	6,0	8,0	10,0
0	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	
1	0,8647	0,9179	0,9502	0,9698	0,9817	0,9889	0,9933	0,9975	0,9997	1,0000
2	,5940	,7127	,8009	,8641	,9084	,9389	,9596	,9826	,9970	0,9995
3	,3233	,4562	,5768	,6792	,7619	,8264	,8753	,9380	,9862	,9972
4	,1429	,2424	,3528	,4634	,5665	,6577	,7350	,8488	,9576	,9897
5	,0527	,1088	,1847	,2746	,3712	,4679	,5595	,7149	,9004	,9707
6	,0166	,0420	,0839	,1424	,2149	,2971	,3840	,5543	,8088	,9329
7	,0045	,0142	,0335	,0653	,1108	,1689	,2378	,3937	,6866	,8699
8	,0011	,0042	,0119	,0267	,0511	,0866	,1334	,2560	,5470	,7798
9	,0002	,0011	,0038	,0099	,0214	,0403	,0681	,1528	,4075	,6672
10	,0000	,0003	,0011	,0033	,0081	,0171	,0318	,0839	,2834	,5421
11		,0001	,0003	,0010	,0028	,0067	,0137	,0426	,1841	,4170
12		,0000	,0001	,0003	,0009	,0024	,0055	,0201	,1119	,3032
13			,0000	,0001	,0003	,0008	,0020	,0088	,0638	,2084
14				,0000	,0001	,0003	,0007	,0036	,0342	,1355
15					,0000	,0001	,0002	,0014	,0173	,0835
16						,0000	,0001	,0005	,0082	,0487
17							,0000	,0002	,0037	,0270
18								,0001	,0016	,0143
19								,0000	,0007	,0072
20									,0003	,0035
21									,0001	,0016
22									,0000	,0007
23										,0003
24										,0001
25										,0000

Tablica 4

Dystrybuanta $\Phi(x)$

x	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	5000	5040	5080	5120	5160	5199	5239	5279	5319	5359
0,1	5398	5438	5478	5517	5557	5596	5636	5675	5714	5753
0,2	5793	5838	5871	5910	5948	5987	6026	6064	6103	6141
0,3	6179	6217	6255	6293	6331	6368	6406	6443	6480	6517
0,4	6554	6591	6628	6664	6700	6736	6772	6808	6844	6879
0,5	6915	6950	6985	7019	7054	7088	7123	7157	7190	7224
0,6	7257	7291	7324	7357	7389	7422	7454	7486	7517	7549
0,7	7580	7611	7642	7673	7704	7734	7764	7794	7823	7852
0,8	7881	7910	7939	7967	7995	8023	8051	8078	8106	8133
0,9	8159	8186	8212	8238	8264	8289	8315	8340	8365	8389
1,0	8413	8438	8461	8485	8508	8531	8554	8577	8599	8621
1,1	8643	8665	8686	8708	8729	8749	8770	8790	8810	8830
1,2	8849	8869	8888	8907	8925	8944	8962	8980	8997	9015
1,3	9032	9049	9066	9082	9099	9115	9131	9147	9162	9177
1,4	9192	9207	9222	9236	9251	9265	9279	9292	9306	9319

x	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
1,5	9332	9345	9357	9370	9382	9394	9406	9418	9429	9441
1,6	9452	9463	9474	9484	9495	9505	9515	9525	9535	9545
1,7	9554	9564	9573	9582	9591	9599	9608	9616	9625	9633
1,8	9641	9649	9656	9664	9671	9678	9686	9693	9699	9706
1,9	9713	9719	9726	9732	9738	9744	9750	9756	9761	9767
2,0	9772	9778	9783	9788	9793	9798	9803	9808	9812	9817
2,1	9821	9826	9830	9834	9838	9842	9846	9850	9854	9857
2,2	9861	9864	9864	9871	9875	9878	9881	9884	9887	9890
2,3	9893	9896	9898	9901	9904	9906	9909	9911	9913	9916
2,4	9918	9920	9922	9925	9927	9929	9931	9932	9934	9936
2,5	9938	9940	9941	9943	9945	9946	9948	9949	9951	9952
2,6	9953	9955	9956	9957	9959	9960	9961	9962	9963	9964
2,7	9965	9966	9967	9968	9969	9970	9971	9972	9973	9974
2,8	9974	9975	9976	9977	9977	9978	9979	9979	9980	9981
2,9	9981	9982	9982	9983	9984	9984	9985	9985	9986	9986
3,0	9987	9987	9987	9988	9988	9989	9989	9989	9990	9990
3,1	9990	9991	9991	9991	9992	9992	9992	9992	9993	9993
3,2	9993	9993	9994	9994	9994	9994	9994	9995	9995	9995
3,3	9995	9995	9995	9996	9996	9996	9996	9996	9996	9997
3,4	9997	9997	9997	9997	9997	9997	9997	9997	9997	9998

x	4,0	4,5	5,0
$\Phi(x)$	$0,9^{4}6833$	$0,9^{5}6602$	$0,9^{6}7134$

Tablica 5

Kwantyle rozkładu normalnego $N(0, 1)$

α	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,5	0,000	0,025	0,050	0,075	0,100	0,126	0,151	0,176	0,202	0,228
0,6	0,253	0,279	0,305	0,332	0,358	0,385	0,412	0,440	0,468	0,496
0,7	0,524	0,553	0,583	0,613	0,643	0,674	0,706	0,739	0,772	0,806
0,8	0,842	0,878	0,915	0,954	0,994	1,036	1,080	1,126	1,175	1,227
0,9	1,282	1,341	1,405	1,476	1,555	1,645	1,751	1,881	2,054	2,326

α	x_α
0,995	2,576
0,999	3,090
0,9995	3,291
0,9999	3,719
0,99995	3,891
0,99999	4,265
0,999995	4,417
0,999999	4,753
0,9999995	4,892
0,9999999	5,199

Tablica 6

Wartości krytyczne $\chi^2(\alpha, n)$ w rozkładzie chi-kwadrat

n \ α	0,995	0,990	0,975	0,950	0,900	0,100	0,050	0,025	0,010	0,005
1	$0,0^4393$	$0,0^3157$	$0,0^3982$	$0,0^2393$	0,0158	2,706	3,841	5,024	6,635	7,879
2	0,0100	0,201	0,0506	0,103	0,211	4,605	5,991	7,378	9,210	10,596
3	0,0717	0,115	0,216	0,352	0,584	6,251	7,815	9,348	11,345	12,838
4	0,207	0,297	0,484	0,711	1,064	7,779	9,488	11,143	13,277	14,860
5	0,412	0,554	0,831	1,145	1,610	9,236	11,070	12,832	15,086	16,750
6	0,676	0,872	1,237	1,635	2,204	10,645	12,592	14,449	16,812	18,548
7	0,989	1,239	1,690	2,167	2,833	12,017	14,067	16,013	18,475	20,278
8	1,344	1,646	2,180	2,733	3,490	13,362	15,507	17,535	20,090	21,955
9	1,735	2,088	2,700	3,325	4,168	14,684	16,919	19,023	21,666	23,589
10	2,156	2,558	3,247	3,940	4,865	15,987	18,307	20,483	23,209	25,188
11	2,603	3,053	3,816	4,575	5,578	17,275	19,675	21,920	24,725	26,757
12	3,074	3,571	4,404	5,226	6,304	18,549	21,026	23,336	26,217	28,300
13	3,565	4,107	5,009	5,892	7,042	19,812	22,362	24,735	27,688	29,819
14	4,075	4,660	5,629	6,571	7,790	21,064	23,685	26,119	29,141	31,319
15	4,601	5,229	6,262	7,261	8,547	22,307	24,996	27,488	30,578	32,801
16	5,142	5,812	6,908	7,962	9,312	23,542	26,296	28,845	32,000	34,267
17	5,697	6,408	7,564	8,672	10,085	24,769	27,587	30,191	33,409	35,718
18	6,265	7,015	8,231	9,390	10,865	25,989	28,869	31,526	34,805	37,156
19	6,844	7,633	8,907	10,117	11,651	27,204	30,144	32,852	36,191	38,582
20	7,434	8,260	9,591	10,851	12,443	28,412	31,410	34,170	37,566	39,997
21	8,034	8,897	10,283	11,591	13,240	29,615	32,671	35,479	38,932	41,401
22	8,643	9,542	10,982	12,338	14,041	30,813	33,924	36,781	40,289	42,796
23	9,260	10,196	11,688	13,091	14,848	32,007	35,172	38,076	41,638	44,181
24	9,886	10,856	12,401	13,848	15,659	33,196	36,415	39,364	42,980	45,558
25	10,520	11,524	13,120	14,611	16,473	34,382	37,652	40,646	44,314	46,928

n \ α	0,995	0,990	0,975	0,950	0,900	0,100	0,050	0,025	0,010	0,005
26	11,160	12,198	13,844	15,379	17,292	35,563	38,885	41,923	45,642	48,290
27	11,808	12,879	14,573	16,151	18,114	36,741	40,113	43,194	46,963	49,645
28	12,461	13,565	15,308	16,928	18,939	37,916	41,337	44,461	48,278	50,993
29	13,121	14,256	16,047	17,708	19,768	39,087	42,557	45,722	49,588	52,336
30	13,787	14,953	16,791	18,493	20,599	40,256	43,773	46,979	50,892	53,672
35	17,192	18,509	20,569	22,465	24,797	46,059	49,802	53,203	57,342	60,275
40	20,707	22,164	24,433	26,509	29,051	51,805	55,758	59,342	63,691	66,766
50	27,991	29,707	32,357	34,764	37,689	63,167	67,505	71,420	76,154	79,490
60	35,535	37,485	40,482	43,188	46,459	74,397	79,082	83,298	88,379	91,952
80	51,172	53,540	57,153	60,391	64,278	96,578	101,879	106,629	112,329	116,321
100	67,328	70,075	74,222	77,929	82,358	118,498	124,342	129,561	135,807	140,169

Tablica 7

Wartości krytyczne $t(\alpha, n)$ w rozkładzie t-Studenta

n \ α	0,20	0,10	0,05	0,02	0,01
1	3,078	6,314	12,706	31,821	63,657
2	1,886	2,920	4,303	6,965	9,925
3	1,638	2,353	3,182	4,541	5,841
4	1,533	2,132	2,776	3,747	4,604
5	1,476	2,015	2,571	3,365	4,032
6	1,440	1,943	2,447	3,143	3,707
7	1,415	1,895	2,365	2,998	3,499
8	1,397	1,860	2,306	2,896	3,355
9	1,383	1,833	2,262	2,821	3,250
10	1,372	1,812	2,228	2,764	3,169
11	1,363	1,796	2,201	2,718	3,106
12	1,356	1,782	2,179	2,681	3,055
13	1,350	1,771	2,160	2,650	3,012
14	1,345	1,761	2,145	2,624	2,977
15	1,341	1,753	2,131	2,602	2,947
16	1,337	1,746	2,120	2,583	2,921
17	1,333	1,740	2,110	2,567	2,898
18	1,330	1,734	2,101	2,552	2,878
19	1,328	1,729	2,093	2,539	2,861
20	1,325	1,725	2,086	2,528	2,845
21	1,323	1,721	2,080	2,518	2,831
22	1,321	1,717	2,074	2,508	2,819
23	1,319	1,714	2,069	2,500	2,807
24	1,318	1,711	2,064	2,492	2,797
25	1,316	1,708	2,060	2,485	2,787

n \ α	0,20	0,10	0,05	0,02	0,01
26	1,315	1,706	2,056	2,479	2,779
27	1,314	1,703	2,052	2,473	2,771
28	1,313	1,701	2,048	2,467	2,763
29	1,311	1,699	2,045	2,462	2,756
30	1,310	1,697	2,042	2,457	2,750
40	1,303	1,684	2,021	2,423	2,704
50	1,299	1,676	2,009	2,403	2,678
60	1,296	1,671	2,000	2,390	2,660
70	1,294	1,667	1,994	2,381	2,648
80	1,292	1,664	1,990	2,374	2,639
90	1,291	1,662	1,987	2,368	2,632
100	1,290	1,660	1,984	2,364	2,626
∞	1,282	1,645	1,960	2,326	2,576

Wartości krytyczne $\begin{matrix} F(0{,}05;\, n, m) \\ F(0{,}01;\, n, m) \end{matrix}$ w rozkładzie F-Snedecora

k \ n	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161,5	199,5	215,7	224,6	230,2	234,0	236,8	238,9	240,5	241,9	243,9	246,0	248,0	249,1	250,1	251,1	252,2	253,3	254,3
	4052	5000	5403	5625	5763	5859	5928	5981	5023	6056	6106	6157	6209	6235	6261	6287	6313	6339	6366
2	18,52	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,39	19,40	19,41	19,43	19,45	19,45	19,46	19,47	19,48	19,49	19,50
	98,50	99,00	99,17	99,25	99,30	99,33	99,36	99,37	99,39	99,40	99,42	99,43	99,45	99,46	99,47	99,47	99,48	99,49	99,50
3	10,13	9,552	9,277	9,117	9,014	8,941	8,887	8,845	8,812	8,786	8,745	8,703	8,660	8,639	8,617	8,594	8,572	8,549	8,527
	43,12	30,82	29,46	28,71	28,24	27,91	27,67	27,49	27,35	27,23	27,05	26,87	26,69	26,60	26,51	26,41	26,31	26,22	26,13
4	7,709	6,944	6,591	6,388	6,256	6,163	6,094	6,041	5,999	5,964	5,912	5,858	5,803	5,774	5,746	5,717	5,688	5,658	5,628
	21,20	18,00	16,69	15,98	15,22	15,21	14,98	14,80	14,66	14,55	14,37	14,20	14,02	13,93	13,84	13,75	13,65	13,56	13,46
5	6,608	5,786	5,410	5,192	5,050	4,950	4,876	4,818	4,773	4,735	4,676	4,619	4,558	4,527	4,496	4,464	4,431	4,398	4,365
	16,26	13,27	12,06	11,39	10,97	10,67	10,46	10,29	10,16	10,05	9,888	9,722	9,553	9,467	9,379	9,291	9,202	9,112	9,020
6	5,987	5,143	4,757	4,534	4,387	4,284	4,207	4,147	4,099	4,060	4,000	3,938	3,874	3,842	3,808	3,774	3,740	3,705	3,669
	13,75	10,93	9,780	9,148	8,746	8,488	8,260	8,102	7,976	7,874	7,718	7,559	7,396	7,313	7,229	7,143	7,057	6,969	6,880
7	5,591	4,737	4,347	4,120	3,973	3,866	3,787	3,726	3,677	3,637	3,575	3,511	3,445	3,411	3,376	3,340	3,304	3,267	3,230
	12,25	9,547	8,451	7,847	7,460	7,191	6,993	6,840	6,719	6,620	6,469	6,314	6,155	6,074	5,992	5,908	5,824	5,737	5,650
8	5,318	4,459	4,066	3,838	3,688	3,581	3,501	3,438	3,388	3,347	3,284	3,218	3,150	3,115	3,079	3,043	3,005	2,967	2,928
	11,26	8,649	7,591	7,006	6,632	6,371	6,178	6,029	5,911	5,814	5,667	5,515	5,359	5,279	5,198	5,116	5,032	4,946	4,859
9	5,117	4,257	3,863	3,633	3,482	3,374	3,293	3,230	3,179	3,137	3,073	3,006	2,937	2,901	2,864	2,826	2,787	2,748	2,707
	11,26	8,022	6,992	6,422	6,057	5,802	5,613	5,467	5,351	5,257	5,111	4,962	4,808	4,729	4,649	4,567	4,483	4,398	4,311
10	4,965	4,103	3,708	3,478	3,326	3,217	3,136	3,072	3,020	2,978	2,913	2,845	2,774	2,737	2,700	2,661	2,621	2,580	2,538
	10,04	7,559	6,552	5,994	5,636	5,386	5,200	5,057	4,942	4,850	4,706	4,558	4,405	4,327	4,247	4,165	4,082	3,997	3,909

k \ n	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
12	4,747	3,885	3,490	3,259	3,106	2,996	2,913	2,849	2,796	2,753	2,687	2,617	2,544	2,506	2,466	2,426	2,384	2,341	2,296
	9,330	6,927	5,953	5,412	5,064	4,821	4,640	4,499	4,386	4,296	4,155	4,010	3,858	3,781	3,701	3,619	3,536	3,44	3,36
15	4,543	3,682	3,287	3,056	2,901	2,791	2,707	2,641	2,588	2,544	2,475	2,404	2,328	2,288	2,247	2,204	2,160	2,114	2,066
	8,683	6,359	5,417	4,893	4,556	4,318	4,142	4,005	3,895	3,805	3,666	3,522	3,372	3,294	3,214	3,132	3,047	2,960	2,868
20	4,351	3,493	3,098	2,866	2,711	2,599	2,514	2,447	2,393	2,348	2,278	2,203	2,124	2,083	2,039	1,994	1,946	1,896	1,843
	8,096	5,849	4,938	4,431	4,103	3,872	3,699	3,564	3,457	3,368	3,231	3,088	2,938	2,859	2,779	2,695	2,608	2,517	2,421
24	4,260	3,403	3,009	2,776	2,621	2,508	2,427	2,355	2,300	2,255	2,183	2,108	2,027	1,984	1,939	1,892	1,842	1,790	1,733
	7,823	5,614	4,718	4,218	3,895	3,667	3,460	3,363	3,256	3,168	3,032	2,889	2,738	2,659	2,577	2,492	2,404	2,310	2,211
30	4,171	3,316	2,922	2,690	2,534	2,421	2,334	2,266	2,211	2,165	2,092	2,015	1,932	1,887	1,841	1,792	1,740	1,684	1,622
	7,563	5,390	4,510	4,018	3,699	3,474	3,305	3,173	3,067	2,979	2,843	2,700	2,549	2,469	2,386	2,299	2,208	2,111	2,006
40	4,085	3,232	2,839	2,606	2,446	2,336	2,249	2,180	2,124	2,077	2,004	1,925	1,839	1,793	1,744	1,693	1,637	1,577	1,509
	7,314	4,977	4,313	3,828	3,514	3,291	3,124	2,993	2,888	2,801	2,665	2,522	2,369	2,288	2,203	2,114	2,019	1,917	1,805
60	4,001	3,150	2,758	2,525	2,368	2,254	2,167	2,097	2,040	1,993	1,917	1,836	1,748	1,700	1,649	1,594	1,534	1,467	1,389
	7,077	4,977	4,126	3,649	3,339	3,119	2,953	2,823	2,716	2,632	2,496	2,352	2,198	2,115	2,029	1,936	1,836	1,726	1,601
120	3,920	3,072	2,680	2,447	2,290	2,175	2,087	2,016	1,959	1,911	1,834	1,751	2,659	1,608	1,554	1,495	1,429	1,352	1,254
	6,851	4,787	3,949	3,480	3,174	2,956	2,792	2,663	2,559	2,472	2,336	2,192	2,035	1,950	1,860	1,763	1,656	1,533	1,381
∞	3,842	2,996	2,605	2,372	2,214	2,099	2,010	1,938	1,880	1,831	1,752	1,666	1,571	1,517	1,459	1,394	1,318	1,221	1,000
	6,635	4,605	3,782	3,319	3,017	2,802	2,639	2,511	2,407	2,321	2,185	2,039	1,878	1,791	1,696	1,592	1,473	1,325	1,000

Wartości krtytyczne $r(\alpha, n)$ współczynnika korelacji

n \ α	0,10	0,05	0,025	0,01	0,005
3	0,9511	0,9877	0,9969	0,9995	0,9999
4	,8000	,9000	,9500	,9800	,9900
5	,6870	,8054	,8783	,9343	,9587
6	0,6084	0,7293	0,8114	,08822	0,9172
7	,5509	,6694	,7545	,8329	,8745
8	,5067	,6215	,7067	,7887	,8343
9	,4716	,5822	,6664	,7498	,7977
10	,4428	,5493	,6319	,7155	,7646
11	0,4187	0,5214	0,6021	0,6851	0,7348
12	,3981	,4973	,5760	,6581	,7079
13	,3802	,4762	,5529	,6339	,6835
14	,3646	,4575	,5324	,6120	,6614
15	,3507	,4409	,5140	,5923	,6411
16	0,3383	0,4259	0,4973	0,5742	0,6226
17	,3271	,4124	,4822	,5577	,6055
18	,3170	,4000	,4683	,5426	,5897
19	,3077	,3887	,4555	52,85	,5751
20	,2992	,3783	,4438	,5155	,5614
21	0,2914	0,3687	0,4329	0,5034	0,5487
22	,2841	,3598	,4227	,4921	,5368
23	,2774	,3515	,4132	,4815	,5256
24	,2711	,3438	,4044	,4716	,5151
25	,2653	,3365	,3961	,4622	,5052
30	0,2407	0,3061	0,3610	0,4226	0,4629
35	,2220	,2826	,3338	,3916	,4296
40	,2070	,2638	,3120	,3665	,4026
45	,1947	,2483	,2940	,3457	,3801
50	,1843	,2353	,2787	,3281	,3610
60	0,1678	0,2144	0,2542	0,2997	0,3301
70	,1550	,1982	,2352	,2776	,3060
80	,1448	,1852	,2199	,2597	,2864
90	,1364	,1745	,2072	,2449	,2702
100	,1292	,1654	,1966	,2324	,2565

Wartości krytyczne $r(\alpha; k, \nu)$ współczynnika korelacji wielokrotnej

	$\alpha = 0{,}05$				$\alpha = 0{,}01$			
ν \ k	3	4	5	6	3	4	5	6
1	0,999	0,999	0,999	0,999	1,000	1,000	1,000	1,000
2	,975	,983	,987	,990	0,995	,0997	0,997	0,998
3	,930	,950	,961	,968	,977	,983	,987	,990
4	,881	,912	,930	,942	,949	,962	,970	,975
5	,836	,874	,898	,914	,917	,937	,949	,957
6	0,795	0,839	0,867	0,886	0,886	0,911	0,927	0,938
7	,758	,807	,838	,860	,855	,885	,904	,918
8	,726	,777	,811	,835	,827	,860	,882	,898
9	,697	,750	,786	,812	,800	,837	,861	,878
10	,671	,726	,763	,790	,776	,814	,840	,859
11	0,648	0,703	0,741	0,770	0,753	0,793	0,821	0,841
12	,627	,683	,722	,751	,732	,773	,802	,824
13	,608	,664	,703	,733	,712	,755	,785	,807
14	,590	,646	,686	,717	,694	,737	,768	,791
15	,574	,630	,670	,701	,677	,721	,752	,776
16	0,559	0,615	0,656	0,687	0,662	0,706	0,738	0,762
17	,545	,601	,641	,673	,647	,691	,724	,749
18	,532	,587	,628	,660	,633	,678	,710	,736
19	,520	,575	,615	,647	,620	,665	,697	,723
20	,509	,563	,604	,636	,607	,652	,685	,712
21	0,498	0,552	0,593	0,624	0,596	0,641	0,674	0,700
22	,488	,542	,582	,614	,585	,630	,663	,690
23	,479	,532	,572	,604	,574	,619	,653	,679
24	,470	,523	,562	,594	,565	,609	,643	,669
25	,462	,514	,553	,585	,555	,600	,633	,660
26	0,454	0,506	0,543	0,576	0,546	0,590	0,624	0,651
27	,446	,498	,536	,568	,538	,582	,615	,642
28	,439	,490	,529	,560	,529	,573	,607	,633
29	,432	,482	,521	,552	,522	,565	,598	,625
30	,425	,476	,514	,545	,514	,557	,591	,618
40	0,373	0,419	0,455	0,484	0,454	0,494	0,526	0,552
60	,308	,348	,380	,406	,377	,414	,442	,467
120	,221	,251	,275	,295	,272	,300	,322	,342

Wartości krytyczne $u\alpha$ serii Walda-Wolfowitza $P(U \leq u_\alpha) \leq \alpha = 0{,}05$

n_2 \ n_1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
4	—	—	2																
5	—	2	2	3															
6	—	2	3	3	3														
7	—	2	3	3	4	4													
8	2	2	3	3	4	4	5												
9	2	2	3	4	4	5	5	6											
10	2	3	3	4	5	5	6	6	6										
11	2	3	3	4	5	5	6	6	7	7									
12	2	3	4	4	5	6	6	7	7	8	8								
13	2	3	4	4	5	6	6	7	8	8	9	9							
14	2	3	4	5	5	6	7	7	8	8	9	9	10						
15	2	3	4	5	6	6	7	8	8	9	9	10	10	11					
16	2	3	4	5	6	6	7	8	8	9	10	10	11	11	11				
17	2	3	4	5	6	7	7	8	9	9	10	10	11	11	12	12			
18	2	3	4	5	6	7	8	8	9	10	10	11	11	12	12	13	13		
19	2	3	4	5	6	7	8	8	9	10	10	11	12	12	13	13	14	14	
20	2	3	4	5	6	7	8	9	9	10	11	11	12	12	13	13	14	14	15

Tablica 12

Wartości krytyczne $u\alpha$ testu rang Wilcoxona-Manna-Whith-neya

$$P(U \leq u_\alpha) \leq \alpha = 0{,}05$$

n_1 / n_2	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2	—	—																		
3	—	—	0																	
4	—	—	0	1																
5	—	0	1	2	4															
6	—	0	2	3	5	7														
7	—	0	2	4	6	8	11													
8	—	1	3	5	8	10	13	15												
9	—	1	3	6	9	12	15	18	21											
10	—	1	4	7	11	14	17	20	24	27										
11	—	1	5	8	12	16	19	23	27	31	34									
12	—	2	5	9	13	17	21	26	30	34	38	42								
13	—	2	6	10	15	19	24	28	33	37	42	47	51							
14	—	2	7	11	16	21	26	31	36	41	46	51	56	61						
15	—	3	7	12	18	23	28	33	39	44	50	55	61	66	72					
16	—	3	8	14	19	25	30	36	42	48	54	60	65	71	77	83				
17	—	3	9	15	20	26	33	39	45	51	57	64	70	77	83	89	96			
18	—	4	9	16	22	28	35	41	48	55	61	68	75	82	88	95	102	109		
19	0	4	10	17	23	30	37	44	51	58	65	72	80	87	99	101	109	116	123	
20	0	4	11	18	25	32	39	47	54	62	69	77	84	92	100	107	115	123	130	138

Tablica 13

Wartości krytyczne $D_n(\alpha)$ w teście Kołmogorowa

n \ α	0,01	0,05
1	0,9950	0,9750
2	,9293	,8419
3	,8290	,7076
4	,7343	,6239
5	,6685	,5633
6	0,6166	0,5193
7	,5758	,4834
8	,5418	,4543
9	,5133	,4300
10	,4889	,4093
11	0,4677	0,3912
12	,4491	,3754
13	,4325	,3614
14	,4176	,3489
15	,4042	,3376

cd

n \ α	0,01	0,05
16	0,3920	0,3273
17	,3809	,3180
18	,3706	,3094
19	,3612	,3014
20	,3524	,2941
30	0,2899	0,2417
40	,2521	,2101
50	,2260	,1884
60	,2067	,1723
70	,1918	,1598
80	,1795	,1496
90	,1694	,1412
100	,1608	,1340

Wartości krytyczne $mnD_{m,n}(\alpha)$ w teście Smirnowa ($n \neq m$)

	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
3		—	—	—	—	—	27	30	33	36	39	42	42	45	48	51	54	57	57	60	63	66	69	72	72	75	79	81
4	—		—	24	28	32	36	36	40	44	48	48	52	56	60	60	64	68	72	72	76	80	84	84	88	92	96	96
5	15	20		30	35	35	40	45	45	50	52	56	60	64	68	70	71	80	80	83	87	90	95	99	100	102	106	110
6	18	20	24		36	40	45	48	54	60	60	64	69	72	73	84	83	88	90	92	97	102	107	108	111	116	116	126
7	21	24	28	30		48	49	53	59	60	65	77	75	77	84	87	91	93	105	103	108	112	115	119	121	133	132	136
8	21	28	30	34	40		55	60	64	68	72	76	81	88	88	94	98	104	107	112	115	128	125	128	135	140	142	148
9	24	28	35	39	42	46		63	70	75	78	84	90	94	99	108	107	111	117	122	126	132	135	138	153	150	153	159
10	27	30	40	40	46	48	53		77	80	84	90	100	100	106	108	113	130	126	130	137	140	150	150	156	158	163	180
11	30	33	39	43	48	53	59	60		86	91	96	102	106	110	118	122	127	134	143	142	150	154	160	162	170	174	180
12	30	36	43	48	53	60	63	66	72		95	104	108	116	119	126	130	140	141	148	149	168	165	168	177	180	187	192
13	33	39	45	52	56	62	65	70	75	81		104	115	121	127	131	138	143	150	156	161	166	172	195	181	191	195	201
14	36	42	46	54	63	64	70	74	82	86	89		123	126	134	140	148	152	161	164	170	176	182	188	189	210	207	210
15	36	44	55	57	62	67	75	80	84	93	96	98		133	142	147	152	160	168	173	179	186	195	199	204	209	215	240
16	39	48	54	60	64	80	78	84	89	96	101	106	114		143	154	160	168	173	180	187	200	199	206	213	220	224	232
17	42	48	55	62	68	77	81	89	93	100	105	111	116	124		164	166	175	180	187	196	203	207	212	221	228	236	242
18	45	50	60	72	72	80	90	92	97	108	110	116	123	128	133		176	182	189	196	204	216	216	224	234	238	244	258
19	45	53	61	70	76	82	89	94	102	108	114	121	127	133	141	142		187	199	204	209	218	224	233	239	248	254	262
20	48	60	65	72	79	88	93	110	107	116	120	126	135	140	146	152	160		199	212	219	228	235	242	250	260	264	280
21	51	59	69	75	91	89	99	105	112	120	126	140	138	145	151	159	163	173		223	227	237	244	249	258	273	273	282
22	51	62	70	78	84	94	101	108	121	124	130	138	144	150	157	164	169	176	183		237	242	250	260	268	274	282	290
23	54	64	72	80	89	98	106	114	119	125	135	142	149	157	163	170	177	184	189	194		249	262	267	275	284	292	298
24	57	68	76	90	92	104	111	118	124	144	140	146	156	168	168	180	183	192	198	204	205		262	276	285	296	301	312
25	60	68	80	88	97	104	114	125	129	138	145	150	160	167	173	180	187	200	202	209	216	225		289	292	300	310	320
26	60	75	80	90	100	108	118	126	132	142	156	156	163	172	178	186	195	202	208	216	223	230	240		304	308	319	328
27	63	73	85	96	101	112	126	130	135	147	154	160	171	176	185	198	201	206	216	222	229	237	243	252		318	329	339
28	66	80	87	98	112	116	125	134	142	152	159	182	175	184	190	198	204	216	224	230	236	244	250	258	264		332	348
29	66	79	87	98	111	120	129	135	145	155	164	168	182	187	197	203	209	219	227	236	241	249	259	267	273	276		346
30	69	82	95	108	113	132	135	150	150	162	167	176	195	197	202	216	217	230	234	242	250	264	270	274	282	288	288	

Wartości krytyczne $nD_{n,n}(\alpha)$ w teście Smirnowe ($n = m$)

n \ α	0,01	0,05	0,10
3	—	—	3
4	—	4	4
5	5	5	4
6	6	5	5
7	6	6	5
8	7	6	5
9	7	6	6
10	8	7	6
11	8	7	6
12	8	7	6
13	9	7	7
14	9	8	7
15	9	8	7
16	10	8	7
17	10	8	8
18	10	9	8
19	10	9	8
20	11	9	8
21	11	9	8
22	11	9	9
23	11	10	9
24	12	10	9
25	12	10	9
26	12	10	9
27	12	10	9
28	13	11	10
29	13	11	10
30	13	11	10
40	15	13	11
50	17	14	13
60	18	15	14
70	20	17	15
80	21	18	16
90	22	19	17
100	23	20	18

Wartości wielomianów ortogonalnych $\xi_p(x)$

a	x	ξ_1	ξ_2	ξ_3	ξ_4	ξ_5
3	1	−1	1			
	2	0	−2			
	3	1	1			
	λ_p	1	3			
4	1	−3	1	−1		
	2	−1	−1	3		
	3	1	−1	−3		
	4	3	1	1		
	λ_p	2	1	$\frac{10}{3}$		
5	1	−2	2	−1	1	
	2	−1	−1	2	−4	
	3	0	−2	0	6	
	4	1	−1	−2	−4	
	5	2	2	1	1	
	λ_p	1	1	$\frac{5}{6}$	$\frac{35}{12}$	
6	1	−5	5	−5	1	−1
	2	−3	−1	7	−3	5
	3	−1	−4	4	2	−10
	4	1	−4	−4	2	10
	5	3	−1	−7	−3	−5
	6	5	5	5	1	1
	λ_p	2	$\frac{3}{2}$	$\frac{5}{3}$	$\frac{7}{12}$	$\frac{21}{10}$
7	1	−3	5	−1	3	−1
	2	−2	0	1	−7	4
	3	−1	−3	1	1	−5
	4	0	−4	0	6	0
	5	1	−3	−1	1	5
	6	2	0	−1	−7	−4
	7	3	5	1	3	1
	λ_p	1	1	$\frac{1}{6}$	$\frac{7}{12}$	$\frac{7}{20}$

a^*	x	ξ_1	ξ_2	ξ_3	ξ_4	ξ_5
8	1	−7	7	−7	7	−7
	2	−5	1	5	−13	23
	3	−3	−3	7	−3	−17
	4	−1	−5	3	9	−15
	λ_p	2	1	$\frac{2}{3}$	$\frac{7}{12}$	$\frac{7}{10}$
9	1	−4	28	−14	14	−4
	2	−3	7	7	−21	11
	3	−2	−8	13	−11	4
	4	−1	−17	9	9	−9
	5	0	−20	0	18	0
	λ_p	1	3	$\frac{5}{6}$	$\frac{7}{12}$	$\frac{3}{20}$
10	1	−9	6	−42	18	−6
	2	−7	2	14	−22	14
	3	−5	−1	35	−17	−1
	4	−3	−3	31	3	−11
	5	−1	−4	12	18	−6
	λ_p	2	$\frac{1}{2}$	$\frac{5}{3}$	$\frac{5}{12}$	$\frac{1}{10}$
11	1	−5	15	−30	6	−3
	2	−4	6	6	−6	6
	3	−3	−1	22	−6	1
	4	−2	−6	23	−1	−4
	5	−1	−9	14	4	−4
	6	0	−10	0	6	0
	λ_p	1	2	$\frac{5}{6}$	$\frac{1}{12}$	$\frac{1}{40}$
12	1	−11	55	−33	33	−33
	2	−9	25	3	−27	57
	3	−7	1	21	−33	21
	4	−5	−17	25	−13	−29
	5	−3	−29	19	12	−44
	6	−1	−35	7	28	−20
	λ_p	2	3	$\frac{2}{3}$	$\frac{7}{23}$	$\frac{3}{20}$

* Dla liczby punktów $a \geq 8$ podano tylko połowę wartości wielomianów $\xi(x)$; pozostałe wartości otrzymuje się z warunku symetrii mianowicie dla $x > \frac{1}{2}a$

$$\xi_p(x) = (-1)^p \xi_p(a + 1 - x)$$

np. dla a = 8 mamy $\xi_1(5) = -\xi_1(4) = 1$, $\xi_2(5) = \xi_2(4) = -5$

Tablica 17

Przekształcenie $z = \frac{1}{2} \ln \frac{1+r}{1-r}$

r	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,0000	0,0100	0,0200	0,0300	0,0400	0,0500	0,0601	0,0701	0,802	0,0902
0,1	,1003	,1104	,1206	,1307	,1409	,1511	,1614	,1717	,1820	,1923
0,2	,2027	,2132	,2237	,2342	,2448	,2554	,2661	,2769	,2870	,2986
0,3	,3095	,3205	,3316	,3428	,3541	,3654	,3769	,3884	,4001	,4118
0,4	,4236	,4356	,4477	,4599	,4722	,4847	,4973	,5101	,5230	,5361
0,5	,5493	,5627	,5763	,5901	,6042	,6184	,6328	,6475	,6625	,6777
0,6	,6931	,7089	,7250	,7414	,7582	,7753	,7928	,8107	,8291	,8480
0,7	,8673	,8872	,9076	,9287	,9505	,9730	,9962	1,0203	1,0453	1,0714
0,8	1,0986	1,1270	1,1568	1,1881	1,2212	1,2562	1,2933	1,3331	1,3758	1,4219
0,9	1,4722	1,5275	1,5890	1,6584	1,7380	1,8318	1,9459	2,0923	2,2976	2,6467

Tablica 18

Przekształcenie $r = \frac{e^{2z} - 1}{e^{2z} + 1}$

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,0000	0,0100	0,0200	0,0300	0,0400	0,0500	0,0599	0,0699	0,798	0,0898
0,1	,0997	,1096	,1194	,1293	,1391	,1489	,1586	,1684	,1781	,1877
0,2	,1974	,2070	,2165	,2260	,2355	,2449	,2543	,2636	,2729	,2821
0,3	,2913	,3004	,3095	,3185	,3275	,3364	,3452	,3540	,3627	,3714
0,4	,3800	,3885	,3969	,4053	,4136	,4219	,4301	,4382	,4462	,4542
0,5	0,4621	0,4699	0,4777	0,4854	0,4930	0,5005	0,5080	0,5154	0,5227	0,5299
0,6	,5370	,5441	,5511	,5580	,5649	,5717	,5784	,5850	,5915	,5980
0,7	,6044	,6107	,6169	,6231	,6291	,6351	,6411	,6469	,6527	,6584
0,8	,6640	,6696	,6751	,6805	,6858	,6911	,6963	,7014	,7064	,7114
0,9	,7163	,7211	,7259	,7306	,7352	,7398	,7443	,7487	,7531	,7574
1,0	0,7616	0,7658	0,7699	0,7739	0,7779	0,7818	0,7857	0,7895	0,7932	0,7969
1,1	,8005	,8041	,8076	,8110	,8144	,8178	,8210	,8243	,8275	,8306
1,2	,8337	,8367	,8397	,8426	,8455	,8483	,8511	,8538	,8565	,8591
1,3	,8617	,8643	,8668	,8692	,8717	,8741	,8764	,8787	,8810	,8832
1,4	,8854	,8875	,8896	,8917	,8937	,8957	,8977	,8996	,9015	,9033
1,5	0,9051	0,9069	0,9087	0,9104	0,9121	0,9138	0,9154	0,9170	0,9186	0,9201
1,6	,9217	,9232	,9246	,9261	,9275	,9289	,9302	,9316	,9329	,9341
1,7	,9354	,9366	,9379	,9391	,9402	,9414	,9425	,9436	,9447	,9458
1,8	,94681	,94783	,94884	,94983	,95080	,95175	,95268	,95359	,95449	,95537
1,9	,95624	,95709	,95792	,95873	,95953	,96032	,96109	,96185	,96259	,96331

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
2,0	0,96403	0,96473	0,96541	0,96609	0,96675	0,96739	0,96803	0,96865	0,96929	0,96986
2,1	,97045	,97103	,97159	,97215	,97269	,97323	,97375	,97426	,97477	,97526
2,2	,97574	,97622	,97668	,97714	,97759	,97803	,97846	,97888	,97929	,97970
2,3	,98010	,98049	,98087	,98124	,98161	,98197	,98233	,98267	,98301	,98335
2,4	,98367	,98399	,98431	,98462	,98492	,98522	,98551	,98579	,98607	,98635
2,5	0,98661	0,98688	0,98714	0,98739	0,98764	0,98788	0,98812	0,98835	0,98858	0,98881
2,6	,98903	,98924	,98945	,98966	,98987	,99007	,99026	,99045	,99064	,990883
2,7	,99101	,99118	,99136	,99153	,99170	,99186	,99202	,99218	,99233	,99248
2,8	,99263	,99278	,99292	,99306	,99320	,99333	,99346	,99359	,99372	,99384
2,9	,99396	,99408	,99420	,99431	,99443	,99454	,99464	,99475	,99485	,99495
z	0,0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9
3	0,99505	0,99595	0,99668	0,99728	0,99777	0,99818	0,99851	0,99878	0,99900	0,99918
4	,99933	,99945	,99955	,99963	,99970	,99975	,99980	,99983	,99986	,99989

ERRATA

STRONA	JEST	POWINNO BYĆ
35	$s = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n-1}$	$s^2 = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n-1}$
42	$\bar{x} - \frac{-\alpha\, t_{(n-1)} \cdot s}{\sqrt{n}} < \mu < \bar{x} + \frac{\alpha\, t_{(n-1)} \cdot s}{\sqrt{n}}$	$\bar{x} - \frac{\alpha\, t_{(n-1)} \cdot s}{\sqrt{n}} < \mu < \bar{x} + \frac{\alpha\, t_{(n-1)} \cdot s}{\sqrt{n}}$
43	136+6 mmHg.	136±6 mmHg.
46	$P\left\{p - \alpha^u \sqrt{\frac{p(1-p)}{n}} < \Pi < p + \alpha^u \sqrt{\frac{p(1-p)}{n}}\right\} \approx 1 - \alpha$	$P\left\{p - \alpha^u \sqrt{\frac{p(1-p)}{n}} < \Pi < p + \alpha^u \sqrt{\frac{p(1-p)}{n}}\right\} \approx 1 - \alpha$
46	$\alpha^u \cdot \sqrt{\frac{p(1-p)}{n}} = 0{,}073 \approx 0{,}07$	$\alpha^u \cdot \sqrt{\frac{p(1-p)}{n}} = 0{,}073 \approx 0{,}07$
59	$s_2 = \frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{(n_1 - 1) + (n_2 + 1)}$	$s^2 = \frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{(n_1 - 1) + (n_2 + 1)}$
60	$s_z = \sqrt{\frac{\sum (z_i - \bar{z})^2}{n-1}}$	$s_z = \sqrt{\frac{\sum (z_i - \bar{z})^2}{n-1}}$
60	$u = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$	$u = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$
60	$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$	$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$
60	$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}{\left(\frac{s_1^2}{n_1}\right)^2 \frac{1}{n_1 + 1} + \left(\frac{s_2^2}{n_2}\right)^2 \frac{1}{n_2 + 1}} - 2$	$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left(\frac{s_1^2}{n_1}\right)^2 \frac{1}{n_1 + 1} + \left(\frac{s_2^2}{n_2}\right)^2 \frac{1}{n_2 + 1}} - 2$

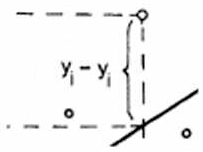
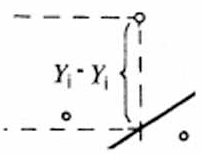
ERRATA

STRONA	JEST	POWINNO BYĆ
61	(2) placedo	(2) placebo
61	(4) placedo	(4) placebo
62	$u = \dfrac{p - \Pi}{\sqrt{\dfrac{\Pi_0 (1 - \Pi_0)}{n}}}$	$u = \dfrac{p - \Pi_0}{\sqrt{\dfrac{\Pi_0 (1 - \Pi_0)}{n}}}$
63	$u = \dfrac{p_1 - p_2}{\sqrt{\dfrac{p (1 - p)}{n}}}$	$u = \dfrac{p_1 - p_2}{\sqrt{\dfrac{p (1 - p)}{n}}}$
68	$t = \dfrac{5{,}23 - 5{,}18}{\sqrt{0{,}002295}}$	$t = \dfrac{5{,}23 - 5{,}18}{\sqrt{0{,}002295}} = 2{,}088$
72	$\chi^2 = \dfrac{(ad - bc)^2}{r_1 r_2 s_1 s_2}$	$\chi^2 = \dfrac{(ad - bc)^2 N}{r_1 r_2 s_1 s_2}$
72	$\chi_c^2 = \sum \dfrac{\left(\lvert ad - bc \rvert - \frac{1}{2} N\right)^2 N}{r_1 r_2 s_1 s_2}$	$\chi_c^2 = \dfrac{\left(\lvert ad - bc \rvert - \frac{1}{2} N\right)^2 N}{r_1 r_2 s_1 s_2}$
78	$r_p = \sqrt{\dfrac{2 (ad - bc)^2}{r_1 r_2 s_1 s_2 + (ad - bc)^2}}$	$r_p = \sqrt{\dfrac{(ad - bc)^2}{r_1 r_2 s_1 s_2 + (ad - bc)^2}}$
78	<table><tr><th></th><th>Tablica</th></tr><tr><td>50 0</td><td>0 50</td></tr><tr><td>50</td><td>50</td></tr><tr><td>40 10</td><td>0 50</td></tr><tr><td>50</td><td>50</td></tr><tr><td>40 40</td><td>10 10</td></tr><tr><td>50</td><td>50</td></tr></table>	<table><tr><th></th><th>Tablica</th></tr><tr><td>50 0</td><td>0 50</td></tr><tr><td>50</td><td>50</td></tr><tr><td>40 10</td><td>0 50</td></tr><tr><td>50</td><td>50</td></tr><tr><td>40 10</td><td>10 40</td></tr><tr><td>50</td><td>50</td></tr></table>

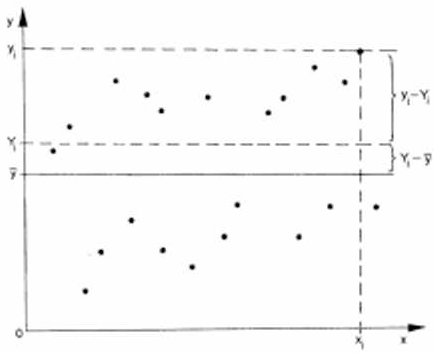
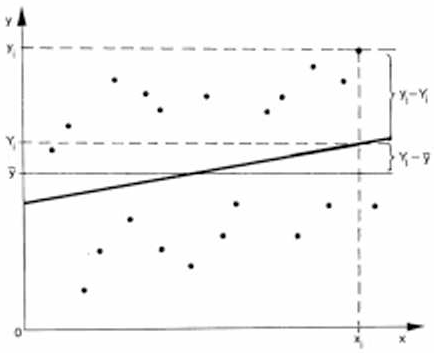
ERRATA

STRONA	JEST	POWINNO BYĆ
80	(Razem)	(Razem) O / E / O – E
80	357 338 19 | 319 338 –10	357 338 19 | 319 338 –19
83	..... $\frac{n_k - r_k}{n_k}$ k r_k p_k	..... $\frac{n_k - r_k}{n_k}$ k wszystkie słupy r_k R N - R n_k N p_k $P = \frac{R}{N}$
85	$\chi_1^2 = \frac{N\left(N\sum_{i=1}^{k} r_i x_i - R\sum_{i=1}^{k} n_i x_i\right)}{R(N-R)\left[N\sum_{i=1}^{k} n_i x_i^2 - \left(\sum_{i=1}^{k} n_i x_i^2\right)\right]}$	$\chi_1^2 = \frac{N\left(N\sum_{i=1}^{k} r_i x_i - R\sum_{i=1}^{k} n_i x_i\right)^2}{R(N-R)\left[N\sum_{i=1}^{k} n_i x_i^2 - \left(\sum_{i=1}^{k} n_i x_i\right)^2\right]}$
93	znajdujemyd χ^2	znajdujemy χ^2
102	$y_{ij} - \bar{y} = (y_{ij} - \bar{y}_i) + (\bar{y}_i - \bar{y})^2$	$y_{ij} - \bar{y} = (y_{ij} - \bar{y}_i) + (\bar{y}_i - \bar{y})$
106	$t = \frac{\bar{y}_g - \bar{y}_h}{s_w \sqrt{\frac{1}{n_g} - \frac{1}{n_h}}}$	$t = \frac{\bar{y}_g - \bar{y}_h}{s_w \sqrt{\frac{1}{n_g} + \frac{1}{n_h}}}$

ERRATA

STRONA	JEST	POWINNO BYĆ
112	$s_i^2 = \dfrac{\sum_{i=1}^{n_i} y_{ij}^2 - \dfrac{\left(\sum_{i=1}^{n_i} y_{ij}\right)^2}{n_i}}{n_i - 1}$	$s_i^2 = \dfrac{\sum_{j=1}^{n_i} y_{ij}^2 - \dfrac{\left(\sum_{i=1}^{n_i} y_{ij}\right)^2}{n_i}}{n_i - 1}$
122	$y_{ij1} \cdot y_{ij2}$ ... $\cdot y_{ijn}$ / T_{ij}	y_{ij1} , y_{ij2} ... , y_{ijn} / T_{ij}
122	$s = \sum_{i,j} y_{ij}^2$	$S = \sum_{i,j,p} y_{ijp}^2$
123	$SKMW = -\dfrac{\sum_i R_i^2}{nc} - \dfrac{T^2}{N}$ $SKMK = -\dfrac{\sum_j C_j^2}{nr} - \dfrac{T^2}{N}$ $SKMW = -\dfrac{\sum_{i,j} T_{ij}^2}{n} - \dfrac{T^2}{N} - SKMW - SKMK$	$SKMW = \dfrac{\sum_i R_i^2}{nc} - \dfrac{T^2}{N}$ $SKMK = \dfrac{\sum_j C_j^2}{nr} - \dfrac{T^2}{N}$ $SKI = \dfrac{\sum_{i,j} T_{ij}^2}{n} - \dfrac{T^2}{N} - SKMW - SKMK$
124	$d_{ij} = y_{ij} - \overline{y}_{i.} - \overline{y}_{.j} + \overline{y}$	$d_{ij} = \overline{y}_{ij} - \overline{y}_{i.} - \overline{y}_{.j} + \overline{y}$
125	$s_{ij} = \sum_{p=1}^{n} y_{ijp}^2$	$S_{ij} = \sum_{p=1}^{n} y_{ijp}^2$
135	*pęcherza (w cm³)*	*pęcherza (w cm²)*
139		

ERRATA

STRONA	JEST	POWINNO BYĆ
145	$Y_0 = \alpha\, t_{(n-2)} \cdot s_0 \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}}$	$Y_0 \pm \alpha\, t_{(n-2)} \cdot s_0 \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}}$
146	Stąd: $a = 2{,}5115$ $b = -0{,}6454$ $r = 0{,}9943$	Stąd: $b = 2{,}5115$ $a = -0{,}6454$ $r = 0{,}9943$
146	$t = r\sqrt{\frac{n-2}{1-r^2}} = 22{,}8296$	$t = r\sqrt{\frac{n-2}{1-r^2}} = 22{,}8296$
149		
150	$s_i = \sum_{j=1}^{n_i} y_{ij}^2 \quad \left\| s_1 \;\; s_2 \;\; s_i \;\; s_k \right\| \quad s = \sum_{i=1}^{k} s_i$	$S_i = \sum_{j=1}^{n_i} y_{ij}^2 \quad \left\| S_1 \;\; S_2 \;\; S_i \;\; S_k \right\| \quad S = \sum_{i=1}^{k} S_i$
157	*trzech pacjentów*	*trzech preparatów*
159	$s^2 = \frac{28{,}1375 + 31{,}3049}{30 + 22 - 4}$	$s^2 = \frac{28{,}1375 + 31{,}3049}{30 + 22 - 4} = 1{,}2384$
159	$3{,}2791 \pm 2{,}013\ \sqrt{0{,}7257} = 4{,}0738$	$3{,}2791 \pm 2{,}013\ \sqrt{0{,}1563} = 4{,}0738$

ERRATA

STRONA	JEST	POWINNO BYĆ
165	$s_C^2 = \frac{85{,}175 + 59{,}4915 - \frac{(17{,}6085 + 8{,}3733)^2}{5{,}4362 + 2{,}4874}}{30 + 22 - 3}$	$s_C^2 = \frac{85{,}175 + 59{,}4915 - \frac{(17{,}6085 + 8{,}3733)^2}{5{,}4362 + 2{,}4874}}{30 + 22 - 3} = 1{,}2137$
170	ocena obiektu	ocena efektu
171	$\sigma^2(M) = \frac{s_c^2}{b^2}\left[\frac{1}{n_1} + \frac{1}{n_2} + \frac{(M - \bar{x}_1 - \bar{x}_2)^2}{(Sx^2)_1 + (Sx^2)_2}\right]$	$\sigma^2(M) = \frac{s_c^2}{b^2}\left[\frac{1}{n_1} + \frac{1}{n_2} + \frac{(M - \bar{x}_1 + \bar{x}_2)^2}{(Sx^2)_1 + (Sx^2)_2}\right]$
172	$\frac{M \pm \frac{\alpha\, t_{(n_1+n_2-3)}\, s_C}{b}\sqrt{(1-g)\cdot\left(\frac{1}{n_1}+\frac{1}{n_2}\right) + \frac{(M-\bar{x}_1-\bar{x}_2)_2}{(Sx^2)_1+(Sx^2)_2}}}{1-g}$	$\frac{M \pm \frac{\alpha\, t_{(n_1+n_2-3)}\, s_C}{b}\sqrt{(1-g)\cdot\left(\frac{1}{n_1}+\frac{1}{n_2}\right) + \frac{(M-\bar{x}_1-\bar{x}_2)^2}{(Sx^2)_1+(Sx^2)_2}}}{1-g}$
173	$0{,}6191 \pm 2{,}102 \cdot \sqrt{0{,}01185} = 0{,}4006$	$0{,}6195 \pm 2{,}102 \cdot \sqrt{0{,}01185} = 0{,}4006$
189	$F_e(x_i) = \frac{\sum_{d \le i} n_j}{N}$	$F_e(x_i) = \frac{\sum_{j \le i} n_j}{N}$