

**ROMUALD STUPNICKI**

**PODSTAWY  
BIOSTATYSTYKI**

*Pamięci moich mistrzów  
Zenona Rogozińskiego  
i Mikołaja Olekiewicza  
poświęcam*

**AKADEMIA WYCHOWANIA FIZYCZNEGO  
JÓZEFA PIŁSUDSKIEGO W WARSZAWIE**

**WARSZAWA 2015**

Copyright © by Romuald Stupnicki

*email:* [rstupnicki@poczta.onet.pl](mailto:rstupnicki@poczta.onet.pl)

Copyright © by Akademia Wychowania Fizycznego  
Józefa Piłsudskiego w Warszawie

*Opracowanie redakcyjne i korekta:* Wiesława Malinowska

ISBN 978-83-61830-02-3

Wydanie I

Wydawnictwo AWF Warszawa  
Warszawa 2015

Objętość 99 stron

Format B-5

# SPIS TREŚCI

Przedmowa	
<b>1. Wprowadzenie</b>	5
<b>2. Zbieranie danych</b>	7
Próby reprezentatywne	7
Dane pomiarowe	8
Dane kwestionariuszowe	8
Rozkłady i ich charakterystyki	9
Badanie rozkładu danych	14
<b>3. Jak poprawnie opisać dane</b>	17
Podstawowe miary zbioru danych	17
Normalizacja danych (transformacje)	20
Transformacja logarytmiczna	20
Transformacja pierwiastkowa	21
Transformacja zmiennych o skali zamkniętej	22
Zmienna unormowana	24
<b>4. Jak zaplanować doświadczenie i ocenić wyniki</b>	26
Porównanie dwu grup	26
Test $t$ Studenta dla grup niezależnych	26
Test $t$ Studenta dla grup zależnych	27
Analiza wariancji (ANOVA)	28
Jednokierunkowa analiza wariancji	28
Układ <i>cross-over</i>	31
Układ krzyżowy	33
Układ bloków losowych	35
Układ hierarchiczno-krzyżowy	36
Jak wstawić brakujące dane	38
Wymagana liczebność grupy	39
Odrzucanie wątpliwych danych	39
Zasady testów statystycznych	40
<b>5. Współzależności między cechami</b>	42
Korelacje dwu zmiennych	42
Regresja liniowa dwu zmiennych	46
Analiza regresji liniowej	48
Regresje krzywoliniowe	51
Przekształcenie regresji w postać liniową	53
Regresja paraboliczna drugiego stopnia (kwadratowa)	54
Wyrównywanie średnią ruchomą	57
<b>6. Analiza wielocechowa</b>	59
Profile wielocechowe	59
Wielocechowa analiza wariancji	61
Korelacje i regresje wielu zmiennych	63
Korelacja i regresja 3 zmiennych	64
Korelacje cząstkowe	66
Korelacje wielu zmiennych	68

<b>7. Analiza danych nieparametrycznych</b> .....	69
Tablice dwu- i wielodzielne .....	69
Nieparametryczny test różnicy między dwiema grupami (test topologiczny) ..	71
Korelacja rangowa (Spearmana) .....	72
<b>8. Zasady tworzenia norm i posługiwanie się nimi</b> .....	74
Wartości referencyjne .....	74
Cechy i rozkłady ich wartości .....	75
Ustalenie przedziału referencyjnego .....	75
Posługiwanie się przedziałami referencyjnymi .....	78
<b>9. Komputer - pomoże, czy zastąpi?</b> .....	80
Wprowadzanie i formatowanie danych .....	80
Funkcje statystyczne programu .....	81
Średnia arytmetyczna, odchylenie standardowe .....	81
Badanie rozkładu .....	82
Współczynnik korelacji .....	83
Regresja dwu zmiennych .....	83
Obliczenia niestandardowe .....	83
<b>10. Tablice statystyczne</b> .....	85
1. Wartości rozkładu normalnego .....	86
2. Wartości dystrybuanty rozkładu normalnego .....	87
3. Wartości funkcji <i>t</i> Studenta .....	88
4. Wartości funkcji <i>F</i> Snedecora .....	89
5. Wartości funkcji $\chi^2$ (chi-kwadrat) .....	91
6. Wartości współczynnika korelacji prostej (Pearsona) .....	92
7. Wartości współczynnika korelacji wielokrotnej .....	93
8. Wartości funkcji $G = [n \cdot \ln(n)]$ .....	94
9. Wartości funkcji <i>z</i> .....	95
10. Wartości funkcji $\arcsin\sqrt{\%}$ .....	96
<b>Skorowidz</b> .....	97
<b>Literatura uzupełniająca</b> .....	99

## Przedmowa

Inspiracją do napisania tej książki były zajęcia dydaktyczne ze studentami warszawskiej AWF, na której nie jest to, mówiąc ogólnie, główny obiekt zainteresowań. Tym cenniejsze było zatem zainteresowanie przedmiotem wykazane przez część studentów, a także pracowników Uczelni. Słowo mówione jest jednak dość ulotne, choć łatwiej nim operować niż pisanym, w którym nie można sobie pozwolić na przejęzyczenia, dygresje i nieściśłości.

Podobno każde równanie umieszczone w książce zmniejsza liczbę jej czytelników o połowę. Ponieważ zaś książeczka ta przeznaczona jest głównie dla tych, „co to pierwszy raz”, postanowiłem ograniczyć liczbę wzorów i formułek do najniezbędniejszego minimum w nadziei, że znajdzie się jeszcze ktoś, kto ją przeczyta, a jak dobrze pójdzie, to może i skorzysta.

Ograniczenie liczby wzorów pociąga za sobą zwrócenie większej uwagi na opis różnych sytuacji, w których stosuje się procedury rachunkowe, a także na najczęściej popełniane błędy. To z kolei wymaga podania i omówienia odpowiednich przykładów. Dalszą konsekwencją było świadome ograniczenie zakresu materiału do zagadnień, z którymi student i pracownik naukowy ma najczęściej do czynienia. Mam nadzieję, że ci, którzy nie znajdują tu rozwiązań swoich problemów, będą umieli skorzystać z obszerniejszych podręczników statystyki matematycznej, bądź z pomocy specjalistów.

*Podstawy Biostatystyki* są poprawioną i uzupełnioną wersją wcześniejszej *Biometrii*, której dwa wydania dawno się rozeszły. Zamieszczone tu przykłady obliczeń w arkuszu EXCEL oparte są na wersjach od 2007 wzwyż, ale mogą być stosowane również we wcześniejszych wersjach Microsoft Office, a także w arkuszu kalkulacyjnym Open Office.

*Romuald Stupnicki*

## 1. WPROWADZENIE

Biostatystyka jest nauką o sposobach zbierania i przetwarzania danych dotyczących szeroko pojętych nauk przyrodniczych, planowania obserwacji i doświadczeń, a także o zasadach poprawnego wnioskowania na podstawie uzyskanych wyników. Jest to więc narzędzie niezbędne nie tylko w pracy badawczej, ale i w praktycznej działalności – w zarządzaniu (klasyfikacja informacji, prognozowanie), w produkcji (kontrola jakości, planowanie wielkości produkcji), a także np. w pracy trenerskiej (ocena postępów w treningu, ranking selekcyjny). W dalszych częściach książki zostaną omówione podstawowe procedury rachunku i wnioskowania statystycznego i właściwe ich stosowanie.

Na początek, należy przypomnieć podział zmiennych na **mierzalne** (ciągłe) i **policzalne** (dyskretne). Przykładem pierwszych są wszelkiego rodzaju pomiary, np. długości, natężenia barwy itp., drugich – liczebności, np. liczby pozytywnych odpowiedzi na pytanie ankiety, liczby krwinek w polu widzenia mikroskopu itp. Rozróżnienie to jest istotne, bowiem cechy mierzalne mają inne właściwości niż policzalne, co pociąga za sobą stosowanie innych technik obliczeniowych. Cechy mierzalne mogą być oparte na **skali naturalnej** (np. masa ciała lub temperatura wyrażona w skali Kelvina) lub **sztucznej** (np. przyrosty masy ciała lub temperatura w skali Celsjusza). To rozróżnienie jest również istotne, bowiem wartości wyrażone w skali naturalnej mogą być przedstawione jako **liczby względne** (np. „przyrost masy ciała wyniósł 20%”), czego nie można zrobić w wypadku skali sztucznej (nie można wyrazić względnego przyrostu temperatury w skali Celsjusza). Skala naturalna może być **otwarta** (nieograniczona „w górę”) lub **zamknięta**. Jako przykład tej ostatniej mogą służyć procenty (skala od zera do stu), współczynniki korelacji (skala od  $-1$  do  $1$ ), zmienne cykliczne (np. miary kątowe).

Na przykładach podanych w tabelce należy również przypomnieć pojęcie **cyfr znaczących**, a także **dokładności liczby**, określone liczbą miejsc dziesiętnych (po przecinku).

wartość	liczba cyfr znaczących	wartość	liczba cyfr znaczących	wartość	liczba cyfr znaczących
25	2	25.0	3	25.04	4
0.25	2	0.025	2	0.2005	4

Statystyka, jako sposób porządkowania i przetwarzania obserwacji, oparta jest na matematyce i logice. Dlatego od użytkownika metod statystycznych wymaga się znajomości podstaw matematyki w zakresie szkoły średniej (funkcje potęgowe, logarytmiczne, równania liniowe, pojęcia związane z prawdopodobieństwem), gdyż ułatwi to korzystanie z opisanych tu procedur.

W książce tej omówiono, na możliwie elementarnym poziomie, tylko wybrane działy biostatystyki. Nie poruszono np. podstaw teorii prawdopodobieństwa, rozkładów innych

niż normalny (np. rozkładu dwumianowego, Poissona i wielu innych), aby uniknąć nadmiaru materiału. Na końcu zamieszczono najczęściej używane tablice statystyczne. Nowością, niepodawaną w większości krajowych podręczników, jest wykorzystanie funkcji  $G$  opartej na logarytmach naturalnych zamiast klasycznej postaci funkcji chi-kwadrat; funkcja  $G$  znacznie ułatwia analizę danych nieparametrycznych. Szczególną uwagę zwrócono na warunki stosowalności najprostszych miar i metod – średniej arytmetycznej, współczynnika korelacji, testu  $t$  Studenta w nadziei, że użytkownik tego zbioru przepisów nie będzie popełniał najpowszechniejszych, elementarnych błędów.

## 2. ZBIERANIE DANYCH

Celem zbierania informacji jest zazwyczaj opis zbiorowości ze względu na jedną lub wiele cech na podstawie **reprezentatywnej próby** lub uzyskanie odpowiedzi na konkretne **pytanie badawcze**. W tym ostatnim wypadku często stosuje się eksperyment. Zebrane informacje poddaje się następnie różnym zabiegom obliczeniowym celem uzyskania możliwie związłego i rzetelnego opisu liczbowego, stwierdzenia ewentualnych zależności między badanymi cechami, wreszcie wyciągnięcia wniosków.

Aby można było zastosować rachunek statystyczny, dane muszą być zebrane w sposób **losowy**, co jest również niezbędnym warunkiem reprezentatywności. W złożonych schematach doświadczalnych losowość może być ograniczona jedynie do ostatniego etapu doboru materiału; zostanie to bardziej szczegółowo omówione w odpowiednich rozdziałach.

### Próby reprezentatywne

Populacja ludzka zamieszkująca jakiś obszar, np. Polskę, jest na ogół bardzo niejednorodna pod względem wieku, wykształcenia, aktywności zawodowej, fizycznej, środowiska geograficznego, społecznego i wielu innych cech. Ponieważ celem badań reprezentatywnych jest opis populacji (a więc większej zbiorowości) na podstawie próby, należy najpierw dokładnie zdefiniować populację. Może to być np. populacja młodzieży męskiej z regionu południowo-wschodniego, albo populacja kobiet ze środowiska wielkomiejskiego lub też populacja bardziej zawężona – np. dorośli uprawiający turystykę zagraniczną, albo studenci zaoczeni jakiejś uczelni.

Jeżeli zdefiniowana populacja jest jednorodna (np. poborowi z określonego miasta), można bezpośrednio wylosować próbę (np. 100 osób). W wypadku złożonych populacji należy najpierw dokonać **warstwowania** jedno- lub wielopoziomowego, np. wg płci, w obrębie płci wg wykształcenia, dalej – wg miejsca zamieszkania itp., a dopiero w obrębie każdej warstwy wylosować pewną liczbę osób, proporcjonalnie do przypuszczalnej liczebności tej warstwy.

Inaczej wygląda dobór grupy, jeżeli celem nie jest charakterystyka określonej populacji, lecz stwierdzenie działania jakiegoś czynnika poprzez przeprowadzenie doświadczenia. Dobór osób do tego doświadczenia nie musi, a często nie powinien być losowy, natomiast podział na grupy doświadczalne musi się odbyć metodą losową. Zasady planowania doświadczeń i odpowiednie schematy układów doświadczeń zostaną omówione w rozdziale 4.

Przed rozpoczęciem zbierania danych należy dokładnie zaplanować schemat obserwacji bądź doświadczenia, bowiem bardzo często okazuje się, że zebrane informacje albo nie nadają się do analizy, albo wskutek niewłaściwego zaplanowania nie da się uzyskać odpowiedzi na postawione pytania. Problemy te będą omówione w następnym rozdziale.



Bardzo często pytamy, ile danych należy zebrać, żeby można było zastosować rachunek statystyczny i wyciągnąć wnioski?. Można odpowiedzieć bardzo ogólnie w ten sposób: im mniejsza jest zmienność między poszczególnymi danymi (np. między osobami), tym mniej obserwacji wystarczy zebrać. Dokładniejsza odpowiedź wymaga znajomości rozrzutu wyników (odchylenia standardowego) w badanym materiale i zostanie omówiona w rozdziale 3.

### **Dane pomiarowe**

Wykonywane pomiary mogą dotyczyć zarówno cech ilościowych, ciągłych, jak i dyskretnych, jakościowych. Przykładem pierwszych mogą być wartości mocy, długości skoku, czasu reakcji itp., drugich – liczba czerwonych krwinek w jednostce objętości, liczba przysiadów w określonym czasie, ale nie liczba przyznanych punktów, bowiem możliwe są wartości ułamkowe. Rozróżnienie to jest istotne, ponieważ cechy ciągle mają najczęściej rozkłady zbliżone do normalnego (po ew. transformacji), natomiast cechy dyskretne mają najczęściej rozkłady typu dwumianowego (patrz dalej).

Pomiary mogą być bezpośrednie, gdy notuje się zmierzoną wartość lub pośrednie, stanowiące np. różnicę między pomiarami bezpośrednimi lub ich iloraz (wartość względną). Należy zwracać uwagę na dokładność zapisu mierzonych wartości: jeżeli np. wiek jest jedną ze zmiennych, a przedział wieku w badanej grupie nie jest duży, wówczas wartość „15 lat” będzie zbyt mało dokładna. Z drugiej strony, wartość masy ciała „54.25 kg” będzie zbyt dokładna (mimo, że taki był wynik pomiaru), gdyż sugeruje „prawdziwość biologiczną”, a przecież masa ciała w danym momencie będzie zależała np. od stanu nawodnienia, treści przewodu pokarmowego itp. Trudno podać prostą regułę rejestracji danych, zależy to bowiem również od typu rozkładu (zob. rozdział 3).

### **Dane kwestionariuszowe**

W wielu wypadkach metoda kwestionariuszowa (ankietowa) jest głównym sposobem zbierania danych, zwłaszcza w badaniach psychologicznych, socjologicznych, pedagogicznych itp., dlatego dalej zostaną omówione zasady tych metod. Wyróżnia się tu dwa główne rodzaje narzędzi: kwestionariusze i ankiety.

*Kwestionariusze* stosowane są często w psychometrii; są to wystandaryzowane zestawy pytań służące do ilościowej oceny jednej lub kilku cech, np. inwentarza osobowości, przy czym dana cecha otrzymuje pewną wartość punktową obliczoną z całego kwestionariusza. Kwestionariusze, zanim staną się standardowym narzędziem, są oceniane pod względem trafności i rzetelności. Dla zwiększenia precyzji często wprowadza się pytania sprawdzające, umieszczone w różnych miejscach kwestionariusza, pozwalające na wyznaczenie tzw. skali kłamstwa. Niekiedy stosuje się tzw. pytania puste, tj. takie, które nie są brane pod uwagę w obliczeniu wartości punktovej. Pytania są typu zero-jedynkowego lub stopniowanego. W tym ostatnim wypadku niektóre pytania są tak sformułowane, że ocena odpowiedzi dokonywana jest w odwrotnej kolejności. Z tych względów do obliczenia punktovej

wartości kwestionariusza służy klucz (szablon), który przykłada się do arkusza celem ułatwienia obliczeń. Skala kwestionariusza jest zamknięta, wartości punktowe zaś można traktować jak cechę ciągłą i poddać parametrycznym metodom obliczeń.

*Ankieta* jest zbiorem pytań dotyczących określonego tematu, np. aktywności fizycznej uczniów, sposobu spędzania wolnego czasu, itp. Ankieta jest zazwyczaj tworzona na różne potrzeby i może zawierać 4 rodzaje pytań:

- wymagające odpowiedzi tak-nie (zero-jedynkowe),
- wymagające tzw. odpowiedzi wielokrotnego wyboru, a więc wskazania jednej lub kilku podanych opcji,
- wymagające wskazania stopnia nasilenia danej cechy (np. czy uprawiasz sportowe gry zespołowe? odpowiedzi: regularnie, dość często, sporadycznie, wcale),
- pytania otwarte, wymagające podania opisu.

W przeciwieństwie do kwestionariuszy, wyniki danej ankiety mogą wymagać stosowania różnych sposobów oceny. Odpowiedzi zero-jedynkowe przedstawia się w postaci procentu odpowiedzi 'tak' (lub 'nie') na dane pytanie, a wyniki analizuje za pomocą funkcji chi-kwadrat. Podobnie postępuje się z odpowiedziami wielokrotnego wyboru, tutaj jednak możliwe jest wskazanie kilku opcji odpowiedzi i dlatego suma odpowiedzi na dane pytanie może być większa niż 100%.

Odpowiedzi stopniowane, wyrażające nasilenie danej cechy, można traktować jak kategorie jakościowe i analizować za pomocą funkcji chi-kwadrat. Jednak stopniowanie tworzy logiczny ciąg ilościowy, dlatego poszczególnym stopniom odpowiedzi można przypisać wartości punktowe, np. 0 ('nie' lub 'wcale'), 1 ('mało' lub 'rzadko') itp., a następnie poddać je procedurom parametrycznym, jak dla cech ciągłych. Skale punktowe powinny być tak dobrane, aby wszystkie odchylenia standardowe były podobne (a nie np. równe pierwiastkowi kwadratowemu ze średniej).

Jeśli idzie o liczbę ankiet (bądź kwestionariuszy) wymaganą do wyciągnięcia miarodajnych wniosków, to trzeba wziąć pod uwagę dwa elementy: zmienność między wynikami oraz „subiektywną rzetelność” danych. O ile błąd pomiarowy wynika z dyskretnego charakteru cech i jest stały, równy pierwiastkowi kwadratowemu z liczby danych, o tyle wpływ subiektywizmu (dającego się niekiedy kontrolować skalą kłamstwa) zawsze będzie zaważył niezbędną liczbę danych. Szczegółowe informacje na temat badań ankietowych zawarte są w książce *Analiza i prezentacja danych ankietowych* (zob. s. 99).

### **Rozkłady i ich charakterystyki**

Przedstawione w poniższej tabelce wyniki skoku w dal pochodzą od uczniów z losowo wybranych klas licealnych. Ponieważ warunki środowiskowo-społeczne w innych szkołach tego typu mogą być różne, wyniki te można traktować jako reprezentatywne dla tej tylko szkoły.

421	446	412	451	502	456	491	516	522	443	390	350	425	475
440	461	486	456	473	473	391	450	511	438	470	381	442	423
413	450	446	416	463	471	410	425	453	426	466	472	450	407
496	502	463	407	461	547	501	482	431	421	392	450	376	412
501	473	463	450	390	330	400	350	410	425	445	483	403	427
390	490	420	430	356	480	430	398	394	425	446	450	372	418
450	475	450	415	355	390	394	415	450	500	401	430	410	416
419	335												

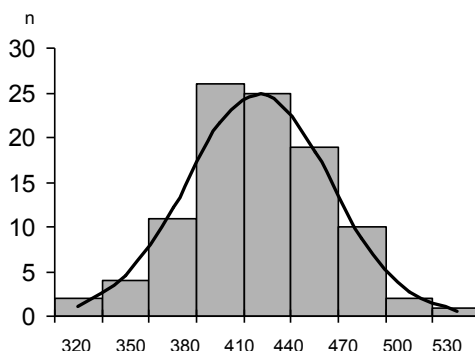
**Tab. 2.1** Wyniki skoku w dal (cm) osiągnięte przez uczniów liceum

Z danych przedstawionych w powyższej tabelce sporządzono **szereg rozdzielczy** (tab. 1.2) przez zliczenie obserwacji w poszczególnych **przedziałach klasowych**. Z szeregu rozdzielczego sporządzono **histogram** (ryc. 1.a), który jest obrazem rozkładu danych. Dla uproszczenia obrazu i ułatwienia obliczeń przedziały klasowe są reprezentowane przez ich środki ( $X$ ).

Przedstawiony rozkład słupkowy charakteryzuje się obecnością jednego szczytu (**wartości najczęstszej** lub **modalnej**) i znaczną symetrią. Ten typ rozkładu nazywa się **rozkładem normalnym**, opisanym równaniem Gaussa. Krzywą rozkładu normalnego pokazano na tym samym wykresie w postaci linii ciągłej. Jak widać, pokrywa się ona zupełnie dobrze z danymi. Badanie zgodności danych z rozkładem normalnym zostanie omówione w dalszej części rozdziału.

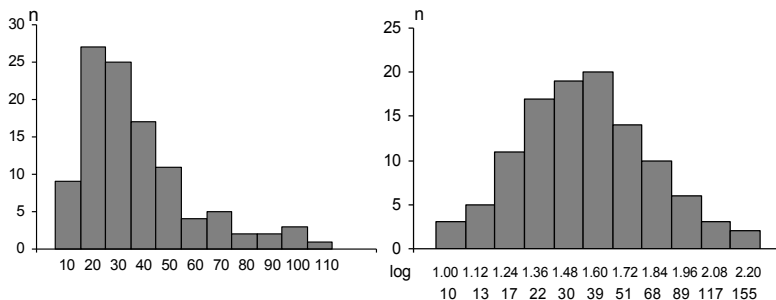
Często zdarza się, że rozkład danych nie przypomina normalnego; na rycinie 2.b pokazano rozkład wartości stężenia ferrytyny w osoczu dziewcząt. Rozkład ten jest wprawdzie jednoszczytowy, ale wyraźnie *prawoskośny* – wartość modalna przesunięta jest silnie w lewo.

Przedziały klasowe	$X$	$n$
305 – 334	320	2
335 – 364	350	4
365 – 394	380	11
395 – 424	410	26
425 – 454	440	25
455 – 484	470	19
485 – 514	500	10
515 – 544	530	2
545 - 584	560	1
	$\Sigma$	100



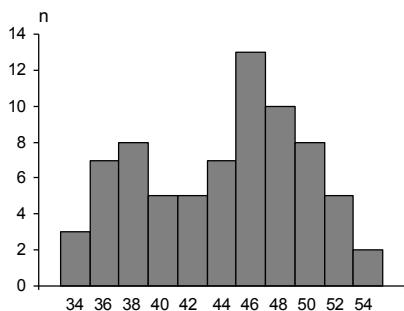
**Tab. 2.2** Szereg rozdzielczy z danych w tabeli 2.1

**Ryc. 2.a** Histogram sporządzony z danych w tabeli 2.2 z nałożoną krzywą rozkładu normalnego



**Ryc. 2.b** Histogramy z szeregu rozdzielczego wartości ferrytyny; z lewej – skala arytmetyczna, z prawej – logarytmiczna

Rozkład prawoskośny można zazwyczaj doprowadzić do postaci normalnej przez zastosowanie innej skali wartości, np. logarytmicznej, jak pokazano na sąsiednim histogramie, a bardziej szczegółowo w następnym rozdziale. Pierwszą wskazówką co do skośności rozkładu może być rozrzut wartości wyrażony współczynnikiem zmienności (patrz s. 19): im większy współczynnik zmienności, tym większa skośność. Gdy współczynnik ten przekracza 20 – 25%, skośność jest zwykle widoczna; jeżeli nie, świadczy to o niejednorodności zbioru danych.



**Ryc. 2.c** Histogram danych o rozkładzie dwuszczytowym

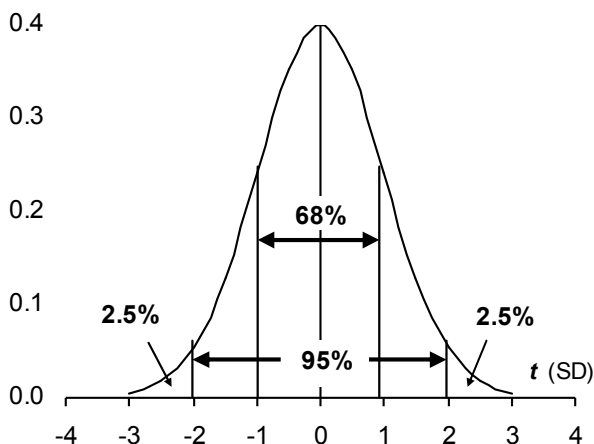
Innym przykładem rozkładu odbiegającego od normalnego jest pokazany na rycinie 2.c rozkład dwuszczytowy (dwumodalny) masy ciała uczniów klas V i VII. Rozkład dwumodalny (lub wielomodalny) świadczy o tym, że badany zbiór danych jest mieszaniną dwu (lub więcej) populacji, jest zatem niejednorodny **statystycznie** (bo niejednorodność widoczna jest na histogramie) i **stochastycznie** (bo jest mało prawdopodobne, by próba wylosowana z jednorodnej populacji wykazywała niejednorodność). Natomiast statystyczna jednorodność rozkładu nie świadczy o jednorodności stochastycznej, bo liczbowo podobne wyniki mogą pochodzić z dwu zupełnie różnych populacji.

Śród wielu rodzajów rozkładów spotykanych w przyrodzie należy tu jeszcze wspomnieć o rozkładzie prostokątnym, charakteryzującym się brakiem wartości modalnej. Roz-

kład taki wystąpi np. dla wieku uczniów szkoły, w której każda klasa ma tyle samo oddziałów. Rozkład taki jest oczywiście nielosowy.

Na rycinie 2.d pokazano krzywą Gaussa, a więc rozkład normalny podobny jak na rycinie 2.a. Można sobie wyobrazić, że powstał on z bardzo dużej, a właściwie nieskończonej liczby danych, z których sporządzono szereg rozdzielnicy o bardzo wielu, bardzo wąskich przedziałach klasowych. Pole pod krzywą Gaussa jest sumą wszystkich słupków histogramu, a więc całkowitą liczebnością zbioru. Liczebność ta jest nieskończenie duża, ale można przyjąć, że całe pole pod krzywą (całkowita liczebność) wynosi 100%. Jeżeli oś  $x$  wyskaluje się nie w jednostkach pomiarowych, lecz w jednostkach odchylenia standardowego od średniej (punkt zero), można obliczyć, jakie będzie pole pod krzywą (procent całkowitej liczebności) w przedziale od minus jednego do plus jednego odchylenia standardowego (średnia  $\pm 1$  SD; posługujemy się tu pojęciami średniej i odchylenia standardowego [SD], które będą omówione na stronie 17).

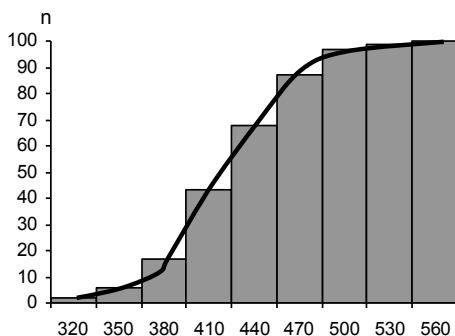
Pole to wynosi ok. 68%, czyli w tym przedziale zawartych jest ok. 68% wszystkich wartości składających się na rozkład normalny. W przedziale: średnia  $\pm 2$  SD znajduje się ok. 95% pola pod krzywą, czyli poza tym przedziałem znajduje się po ok. 2.5% pola z każdej strony, łącznie ok. 5% (patrz wykres).



**Ryc. 2.d** Krzywa rozkładu normalnego (krzywa Gaussa). Wartość zero na osi poziomej odpowiada średniej arytmetycznej.

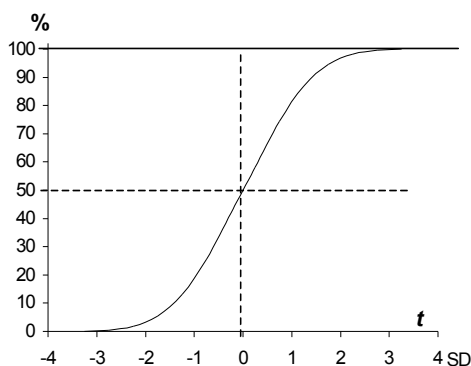
Dane z przykładu w tabeli 2.2 pokazano jeszcze raz w tabeli 2.3 z tym, że oprócz liczebności każdego przedziału klasowego ( $n$ ) podano także **liczebności skumulowane** ( $N$ ; sumowane dla kolejnych przedziałów klasowych). Wykres liczebności skumulowanych pokazano na wykresie słupkowym (ryc. 2.e). Wykres taki nazywa się **dystrybuantą** rozkładu.

X	n	N
320	2	2
350	4	6
380	11	17
410	26	43
440	25	68
470	19	87
500	10	97
530	2	99
560	1	100
$\Sigma$	100	



**Tab. 2.3** Dane z tabeli 1.2 z dodanymi wartościami skumulowanymi

**Ryc. 2.e** Wykres słupkowy danych z tabeli 1.3 z nanesioną linią dystrybuanty



**Ryc. 2.f.** Dystrybuanta rozkładu normalnego

Dystrybuanta rozkładu normalnego ma postać symetrycznej krzywej „S”, jak pokazano na wykresie 2.f. Z dystrybuanty, której wartości podane są w tabelicy 2, można odczytać wielkość pola pod krzywą rozkładu normalnego od początku (minus nieskończoności) do danego punktu (wartości  $t$ , czyli liczby odchyłeń standardowych). W ten sposób można obliczyć, jaki procent populacji znajdzie się w żądanym przedziale. Poniżej podano przykłady takich obliczeń opartych na średnich i SD.

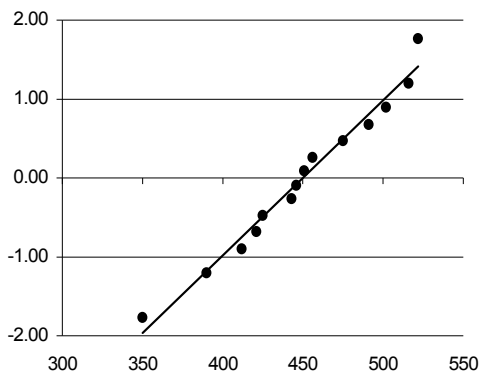
1. Z szeregu rozdzielczego wartości skoku w dal obliczono średnią arytmetyczną (434.4) i SD (45.5) jak na s. 18. Pytanie brzmi: u jakiego odsetka populacji można się spodziewać wyników powyżej 500 cm? Zakładamy, że próba jest reprezentatywna, a rozkład nie odbiega od normalnego. Wartość 500 cm należy wyrazić w jednostkach odchyłeń standardowych od średniej arytmetycznej (wartość unormowana; patrz s. 24):  $t = (500 - 434.3)/45.5 = 1.44$ . W tabelicy 2 znajdujemy wartość dystrybuanty odpowiadającą  $t = 1.44$ ; wynosi ona 0.925; znaczy to, że 92.5% populacji osiąga wyniki skoku do 500 cm, a reszta, czyli 7.5%, powyżej 500 cm.

2. Opierając się na tych samych danych, należy znaleźć procent populacji dla przedziału długości skoku od 350 do 400 cm. Tak jak poprzednio, obliczamy wartości unormowane dla 350 i 400 cm:  $t_{350} = (350 - 434.4)/45.5 = -1.85$  oraz  $t_{400} = (400 - 434.4)/45.5 = -0.76$ . Odpowiadające im wartości dystrybucyjności wynoszą 0.032 i 0.224, czyli 3.2% populacji osiąga wyniki do 350 cm, a 22.4% – do 400 cm. Zatem w przedziale 350–400 cm będzie zawarte  $22.4 - 3.2 = 19.2\%$  populacji. Podobnie jak w poprzednim przykładzie, rozumowanie to będzie poprawne, jeżeli spełnione są dwa założenia: że próba jest reprezentatywna dla danej, odpowiednio zdefiniowanej populacji, i że rozkład danych jest normalny.

### Badanie rozkładu danych

Badanie rozkładu danych, a więc określenie jego charakterystyki, przeprowadza się w celu sprawdzenia, czy spełnione są wymagania dotyczące stosowania testów statystycznych lub opisu zbiorowości. O ile jednak większość testów statystycznych jest względnie mało wrażliwa na odchylenia od normalności rozkładu i wystarcza zastosowanie przybliżonego testu na normalność, o tyle do statystycznego opisu populacji (i w niektórych innych sytuacjach) konieczne jest dokładne zbadanie rozkładu danych. Badanie rozkładu polega na ocenie zgodności z rozkładem normalnym, a więc na ocenie symetrii i jednorodności zbioru danych. Dla potrzeb testów statystycznych, np. analizy wariancji, często stosuje się test Shapiro-Wilka (dla  $n < 50$ ) lub test d'Agostino (dla  $n > 50$ ). Są one dostępne w programach statystycznych i nie będą tu omawiane, gdyż są dość uciążliwe rachunkowo.

412	-1.73
413	-1.16
415	-0.86
417	-0.62
417	-0.37
423	-0.19
425	0.00
429	0.19
430	0.37
431	0.62
434	0.86
437	1.16
439	1.73



**Tab. 2.4** Wyniki skoku w dal i odpowiadające im wartości ON

**Ryc. 2.g** Wykres danych z tabeli 2.4

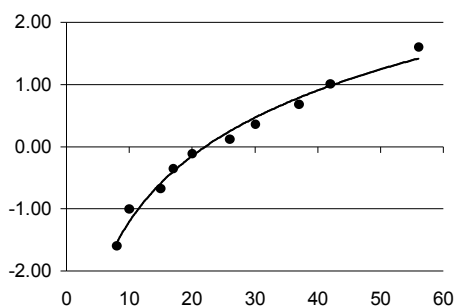
Najprostszym, a jednocześnie najlepszym sposobem jest graficzna ocena rozkładu. Polega ona na naniesieniu na wykres badanych wartości, uporządkowanych od najmniejszej do największej, w skali odpowiadającej rozkładowi normalnemu, co najlepiej można wyko-

nać korzystając z funkcji arkusza EXCEL (zob. s. 81). Przy dużej liczbie danych, rzędu setek, można je również ułożyć w szereg rozdzielczy, przedstawić w postaci histogramu i nanieść krzywą rozkładu normalnego. Sposób ten nie będzie tu jednak omówiony.

Dane zawarte w tabeli 2.4 zawierają uporządkowane wartości z pierwszego wiersza wyników skoku w dal (tab. 2.1), a w sąsiedniej kolumnie podane są wartości „odchyleń normalnych”, czyli unormowanych wartości dystrybuanty rozkładu normalnego dla tej samej liczebności (kolumna D w tabeli na s. 82). Jeżeli badany zbiór ma rozkład normalny, wówczas wykres odchyleń normalnych względem uporządkowanych wartości badanego zbioru będzie liniowy, jak pokazano na rycinie 2.g.

Dane przedstawione w tabeli 2.5 (stężenia ferrytyny w osoczu) są powtórzeniem przy-  
kładu ze s. 20 („Transformacje logarytmiczne”). Jak widać, wykres ma kształt łukowatej linii, co świadczy o prawoskośności. Przykład histogramu takiego rozkładu pokazano na rycinie 2.b.

8	-1.60
10	-1.01
15	-0.68
17	-0.36
20	-0.12
26	0.12
30	0.36
37	0.68
42	1.01
56	1.60

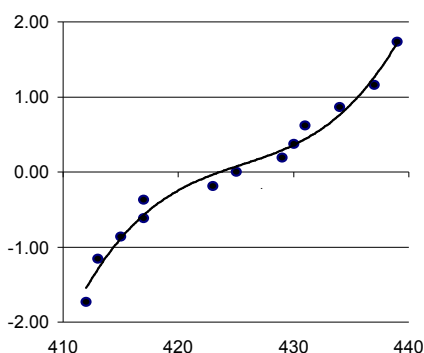


**Tab. 2.5** Stężenia ferrytyny i odpowiadające im wartości ON

412	-1.73
413	-1.16
415	-0.86
417	-0.62
417	-0.37
423	-0.19
425	0.00
429	0.19
430	0.37
431	0.62
434	0.86
437	1.16
439	1.73

**Tab. 2.6** Wyniki skoku w dal i odpowiadające im wartości ON

**Ryc. 2.h** Wykres danych z tabeli 2.5



**Ryc. 2.i** Wykres danych z tabeli 2.6



Wykres pokazany przy tabeli 2.6 świadczy o niejednorodności zbioru danych – wartości powinny być skupione wokół średniej, tu natomiast przeważają wartości niskie bądź wysokie. Jeżeli jest to próba reprezentatywna, to populacja, z której ta próba pochodzi, jest najwyraźniej niejednorodna.

Zbiory danych liczące ok. 100 elementów lub więcej można najpierw ułożyć w szereg rozdzielczy zwykły i skumulowany, a następnie ten ostatni wyrazić w procentach całkowitej liczebności, lub zastosować procedurę opisaną na s. 82 („Badanie rozkładu”).

W tabeli 2.7 pokazano przykład podobny jak w tabeli 2.3, ale o nieco większej liczebności całkowitej. Dane te naniesiono na siatkę normalną (ryc. 2.k) i otrzymano wykres liniowy, co świadczy o normalności rozkładu.

X	n	N	%N
320	2	2	1.9
350	5	7	6.7
380	14	21	20.0
410	26	47	44.8
440	25	72	68.6
470	19	91	86.7
500	10	101	96.2
530	3	104	99.0
560	1	105	100
$\Sigma$	100		

**Tab. 2.7** Szereg rozdzielczy wyników oraz bezwzględne (N) i względne (%N) liczebności skumulowane

W omówionych przykładach stosowano „odchylenia normalne” obliczone jak opisano na s. 82, oparte na tej samej zasadzie co siatka normalna, a więc na dystrybucie rozkładu normalnego.

### 3. JAK POPRAWNIE OPISAĆ DANE

Jak wspomniano w rozdziale 2, sposób opisu danych będzie zależał od sytuacji. Chcąc scharakteryzować populację pod względem jednej lub wielu cech stosuje się nieco inne kryteria, niż gdy opisuje się wyniki doświadczenia. W tym ostatnim wypadku wymóg reprezentatywności jest często zbędny, podczas gdy dla opisu populacji ma on pierwszorzędne znaczenie; podobnie jest z innymi wymogami – jednorodnością i symetrią rozkładu itp. Poniżej zostaną opisane różne miary używane w opisie danych i warunki stosowalności tych miar.

#### Podstawowe miary zbioru danych

Najpełniejszym opisem zbioru danych, tak liczbowych jak i tekstowych, jest sam zbiór, a więc po prostu wszystkie dane. Jeżeli zbiór nie jest liczny, zawiera np. tylko kilka elementów, nie sprawi to większych kłopotów. Jeżeli jednak zbiór jest liczny, a zależy nam jedynie na jego ogólnej charakterystyce, najlepsze będą miary zbiorcze (dla zbioru liczby) lub krótki opis streszczający zbiór tekstowy.

Dla zbiorów liczbowych konieczne są zazwyczaj dwie miary: jedna, będąca jakąś wartością przeciętną i druga, charakteryzująca rozrzut wartości, przy czym należy uwzględnić rozkład zbioru, omówiony w poprzednim rozdziale. Zostaną tu opisane trzy miary przeciętne (inaczej – centralne) zbioru: wartość średnia, wartość środkowa (**mediana**) i wartość występująca najczęściej (**modalna**), a także dwie miary rozrzutu – zakres wartości (**rozstęp**) i **odchylenie standardowe** (a także wielkości pochodne).

O ile pojęcie średniej arytmetycznej jest raczej oczywiste i nie ma potrzeby objaśniania go bliżej, o tyle odchylenie standardowe (*standard deviation*; SD) nie jest miarą „intuicyjną”. Jeżeli np. mamy 5 liczb – 8, 9, 9, 10 i 14, to ich średnia arytmetyczna będzie równa 10. Jeżeli teraz od każdej liczby odejmiemy średnią arytmetyczną, otrzymamy szereg: 8 – 10, 9 – 10, 9 – 10, 10 – 10 i 14 – 10, czyli –2, –1, –1, 0 i 4. Łatwo się przekonać, że suma tych *odchyleń od średniej arytmetycznej* jest równa zero. Następnie, podnosząc poszczególne odchylenia do kwadratu otrzymamy 4, 1, 1, 0 i 16. Suma *kwadratów odchyleń* jest tu równa 22. Następnie obliczamy średni kwadrat odchyleń, ale w tym celu nie dzielimy wartości 22 przez liczbę odchyleń (5), lecz przez tzw. **liczbę stopni swobody** (*degrees of freedom*; df). Liczba stopni swobody jest liczbą tzw. obserwacji niezależnych, co można wyjaśnić następująco: aby otrzymać 22 jako sumę 5 liczb, można cztery liczby wybrać zupełnie dowolnie, natomiast piąta liczba musi być uzupełnieniem sumy tych 4 liczb do wartości 22. Mówimy, że suma pięciu liczb ma cztery stopnie swobody, bardziej ogólnie – że suma  $n$  liczb ma  $n - 1$  stopni swobody. Wracając do sumy kwadratów odchyleń, dzieli się ją przez liczbę stopni swobody, a więc przez 4. Otrzymany średni kwadrat odchyleń  $22/4 = 5.5$  nazywa się **wariancją**, a pierwiastek z wariancji  $\sqrt{5.5} = 2.35$  – **odchyleniem standardowym**.

W praktyce, odchylenia standardowego nie oblicza się w sposób podany powyżej, byłoby to bowiem uciążliwe, zwłaszcza przy większej liczbie danych. Podany niżej sposób jest dużo prostszy rachunkowo, a matematycznie równoważny pokazanej zasadzie.

Jako przykład niech posłużą poniższe dane dotyczące względnej zawartości (“procentu”) tłuszczu w ciele (%F). Dla ułatwienia niektórych obliczeń uporządkowano wartości od najmniejszej do największej.

10, 11, 14, 14, 15, 16, 16, 16, 18, 18, 19, 19, 20, 22, 23.

Średnią ( $\bar{x}$ ) i odchylenie standardowe (s; SD) oblicza się następująco:

$n = 15$ $\sum X_i = 10+11+ \dots +23 = 251$ $\sum X_i^2 = 100+121+ \dots +529 = 4389$	$\bar{x} = \sum X_i/n = 251/15 = 16.73$ $S_{xx} = \sum X_i^2 - (\sum X_i)^2/n = 4389 - 251^2/15 = 188.93$ $SD = \sqrt{\frac{S_{xx}}{n-1}} = \sqrt{188.93/14} = \sqrt{13.50} = 3.67$
--	---

Powyższy przykład ukazuje zasadę obliczeń. W praktyce, jeżeli nie są potrzebne inne rachunki, np. analiza wariancji, a jedynie obliczenie odchylenia standardowego lub średniej, najlepiej skorzystać z funkcji arkusza EXCEL, jak pokazano na s. 82.

Zapis uzyskanych wyników: **średnia  $\pm$ SD**: %F = 16.73  $\pm$  3.67 (n = 15). Inne miary: **błąd średniej (błąd standardowy; SE)**: SE = SD/  $\sqrt{n}$  = 3.67/  $\sqrt{15}$  = 0.95; **zakres wartości**: 10 ÷ 23; **rozstęp** = 23 – 10 = 13; **mediana**: 16 (środkowa, ósma wartość); **modalna**: 16, bo są trzy takie wartości, ale zbiór jest zbyt mały, aby modalna mogła być wyznaczona dokładnie. O ile ustalenie mediany nie budzi wątpliwości, gdyż jest to dokładnie środek uporządkowanego zbioru, o tyle modalnej nie można wyznaczyć bezpośrednio, a oszacować ją można z histogramu, ew. obliczyć z funkcji rozkładu odpowiadającego temu histogramowi. W przybliżeniu, modalna jest równa kwadratowi mediany podzielonemu przez średnią.

Wspomniane trzy miary przeciętne można scharakteryzować następująco:

**Mediana** - miara pozycyjna, niezależna od rozkładu, często wykorzystywana w tzw. testach nieparametrycznych. Dalszymi miarami pozycyjnymi są **kwartyle**, wyznaczające jedną czwartą i trzy czwarte uporządkowanego zbioru.

**Modalna** - jako przedział wartości najczęstszych, a więc najbardziej typowych, jest „naturalną” miarą charakteryzującą populację. Dlatego w opisie populacji zawsze idzie o wyznaczenie wartości modalnej. Ponieważ nie da się tego dokonać bezpośrednio, używa się średniej arytmetycznej jako reprezentacji wartości modalnej. To jest jednak możliwe tylko wówczas, gdy rozkład wartości w populacji jest normalny, wówczas bowiem wszystkie trzy miary – modalna, mediana i średnia arytmetyczna pokrywają się.

**Średnia arytmetyczna** – jest odpowiednikiem środka ciężkości w mechanice, nie ma jednak poza tym interpretacji biologicznej. Jej zaletą jest łatwość wyznaczenia i przydatność w rachunku statystycznym. Należy zawsze pamiętać, że

**średnia arytmetyczna może być sensownie stosowana do opisu zbiorowości tylko wówczas, gdy pokrywa się z wartością modalną.**

Przedstawiony sposób obliczeń może być niepraktyczny, jeżeli danych jest bardzo dużo i nie są wprowadzone do komputera, np. jako zbiór Excela. Poniżej pokazano sposób postępowanie dla większej liczby danych..

Przedz.klas.	X	n	$n \cdot X$	$n \cdot X^2$
305 – 334	320	2	640	204 800
335 – 364	350	4	1 400	490 000
365 – 394	380	11	4 180	1 588 400
395 – 424	410	26	10 660	4 370 600
425 – 454	440	25	11 000	4 840 000
455 – 484	470	19	8 930	4 197 100
485 – 514	500	10	5 000	2 500 000
515 – 544	530	2	1 060	561 800
545 – 584	560	1	560	313 600
$\Sigma$		100	43 430	19 066 300

**Tab. 3.1** Szereg rozdzielczy (powtórzenie tabeli 1.2) zawierający dodatkowo sumy ( $n \cdot X$ ) i sumy kwadratów ( $n \cdot X^2$ ) do obliczenia średniej i odchylenia standardowego

Dla przedziałów klasowych (dane z tab. 2.1 i 2.2) wyznaczono ich środki (X), a następnie obliczono sumy danych ( $n \cdot X$ ) i sumy kwadratów danych ( $n \cdot X^2$ ) w każdej klasie (tab. 3.1). Po zsumowaniu wszystkich klas obliczono średnią arytmetyczną i odchylenie standardowe z szeregu rozdzielczego:

$$\bar{x} = 43\,430/100 = 434.3, \quad S_{xx} = 19\,066\,300 - 43\,430^2/100 = 204\,651, \\ SD = \sqrt{204\,651/99} = 45.47.$$

Ponieważ odchylenie standardowe obliczone z szeregu rozdzielczego jest zazwyczaj przesadnie duże w stosunku do obliczonego z indywidualnych danych, wprowadza się poprawkę Shepparda:

$$s^{2'} = s^2 - 30^2/12,$$

gdzie  $s^2$  jest wariancją, 30 jest szerokością przedziału klasowego, a 12 – stałą. Zatem  $s^{2'} = 2067.2 - 900/12 = 1992.2$ , a  $SD = \sqrt{1992.2} = 44.6$ .

Mimo poprawki, SD obliczone z szeregu rozdzielczego (44.6) jest nadal nieco większe od obliczonego z indywidualnych danych (42.7).

Na zakończenie należy wspomnieć o bardzo przydatnej względnej mierze rozrzutu. Jeżeli odchylenie standardowe podzieli się przez średnią arytmetyczną i pomnoży przez 100, otrzyma się tzw. **współczynnik zmienności** (*coefficient of variability*; CV):

$$CV = 100 \cdot \frac{SD}{\bar{x}} \quad [3-1]$$

Współczynnik zmienności wyrażany jest zatem w „procentach” średniej arytmetycznej i umożliwia porównywanie cech wyrażanych w różnych jednostkach pod względem rozrzutu.

## Normalizacja danych (transformacje)

Często zdarza się, że zbiór wartości ma rozkład wyraźnie odbiegający od normalnego, co może utrudnić lub uniemożliwić opis statystyczny lub przeprowadzenie odpowiednich testów. W takim wypadku należy dane tak przeskalać (przekształcić matematycznie, czyli poddać transformacji), aby uzyskać rozkład zbliżony do normalnego. Najczęstsze sytuacje, kiedy transformacja jest pożądana lub konieczna, są następujące:

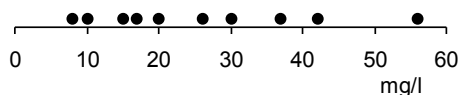
1. Doprowadzenie zbioru danych do postaci zbliżonej do rozkładu normalnego,
2. Wyrównanie wariancji (zmienności) w grupach,
3. Przekształcenie zależności krzywoliniowej w liniową.

Poniżej podano rachunki związane z transformacją zmiennych, natomiast wspomniane zastosowania zostaną omówione w odpowiednich rozdziałach.

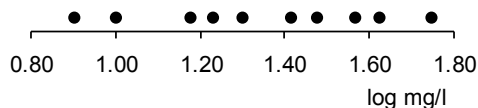
### Transformacja logarytmiczna

W tabeli 3.2 pokazano stężenia ferrytyny ( $x$ ) w osoczu krwi. Łatwo zauważyć, że rozkład wartości jest silnie skośny, bo: niskie wartości są bardziej zagęszczone niż wysokie (ryc. 3.a), współczynnik zmienności jest bardzo duży (55.5%), a porównanie z rozkładem normalnym daje linię krzywą (ryc. 3.c). Po zlogarytmowaniu danych rozkład punktów jest symetryczny (ryc. 3.b), a porównanie z rozkładem normalnym daje linię prostą (ryc. 3.b). Ponadto, dla danych nielogarytmowanych średnia minus 2 odchylenia standardowe daje wartość ujemną (-2.9), co jest oczywistym nonsensem.

$x$	$\log(x)$
8	0.903
10	1.000
15	1.176
17	1.230
20	1.301
26	1.415
30	1.477
37	1.568
42	1.623
56	1.748
$\bar{x}$ 26.1	1.344
SD 14.5	0.259
CV = 55.5%	

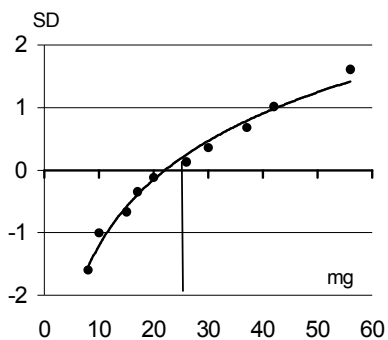


Ryc. 3.a Rozkład wartości ferrytyny ( $x$ )

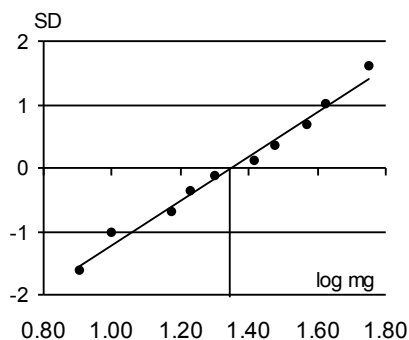


Ryc. 3.b Rozkład logarytmów wartości ferrytyny ( $\log(x)$ )

Tab. 3.2 Stężenia ferrytyny ( $x$ ) w osoczu krwi i logarytmy tych stężeń ( $\log(x)$ )



**Ryc. 3.c** Porównanie wartości ferrytyny (x) z rozkładem normalnym



**Ryc. 3.d** Porównanie logarytmów wartości ferrytyny ( $\log(x)$ ) z rozkładem normalnym

W tabeli 3.3 pokazano wartości średniego logarytmu  $\pm 1$  i 2 SD oraz odpowiadające im wartości zdelogarytmowane. Jak widać, odległości między kolejnymi wartościami zdelogarytmowanymi systematycznie rosną – każda następna jest równa poprzedniej razy 1.945, czyli zdelogarytmowanej wartości odchylenia standardowego (0.259). Średnia wartość logarymiczna ( $\pm$ SD) wynosi tu  $1.344 \pm 0.259$ , co po zdelogarytmowaniu daje  $22.1 \cdot 1.945^{\pm 1}$ . Zapis 1.945 do potęgi  $\pm 1$  oznacza, że średnią geometryczną, czyli zdelogarytmowaną średnią (22.1) należy pomnożyć i podzielić przez zdelogarytmowane odchylenie standardowe (1.945). W ten sposób wartości mniejsze od średniej nigdy nie osiągną zera, a więc opis za pomocą miar statystycznych będzie lepiej odpowiadał rzeczywistości, niż gdyby do powyższych danych zastosowano miary arytmetyczne ( $26.1 \pm 14.5$ ). Należy tu dodać, że średnia geometryczna będzie zawsze mniejsza od arytmetycznej dla tego samego zbioru danych a w rozkładzie prawoskośnym będzie bardziej zbliżona do wartości modalnej niż średnia arytmetyczna.

SD	log c	c
-2	0.826	6.7
-1	1.085	12.2
0	1.344	22.1
1	1.603	40.1
2	1.862	72.8

**Tab. 3.3** Odchylenia logarymicznych wartości ferrytyny ( $\log(x)$ ) od średniej i odpowiadające tym odchyleniom wartości zdelogarytmowane

### Transformacja pierwiastkowa

W niektórych wypadkach odchylenie standardowe równe jest w przybliżeniu pierwiastkowi ze średniej arytmetycznej. Dotyczy to zwłaszcza zmiennej dyskretnej, np. liczby celnych podań na 10 strzałów lub pomiarów czasu, np. biegu (patrz tab. 3.4). Do takich danych

stosuje się przekształcenie pierwiastkowe, które daje podobny efekt jak logarytmiczne, tzn. normalizuje rozkład danych, a średnia pierwiastkowa (kwadrat średniego pierwiastka) jest bardziej zbliżona do wartości modalnej niż średnia arytmetyczna. Podobnie jak średnia geometryczna, również średnia pierwiastkowa jest zawsze mniejsza od arytmetycznej, jeżeli indywidualne wartości są większe od 1. Nie da się tu zastosować opisu podobnego do wartości logarytmicznych i trzeba przedstawić tabelkę danych, jak pokazano obok dla czasu biegu.

liczba podań		czas biegu	
n	$\sqrt{n}$	s	$\sqrt{s}$
2	1.41	77	8.77
3	1.73	80	8.94
3	1.73	81	9.00
3	1.73	82	9.06
4	2.00	83	9.11
4	2.00	85	9.22
5	2.24	87	9.33
7	2.65	87	9.33
8	2.83	89	9.43
9	3.00	90	9.49
		94	9.70
		98	9.90
		112	10.58
średnie i odchylenia standardowe			
4.80	2.13	89.8	9.47
2.27	0.50	8.9	0.46

**Tab. 3.4** Liczby skutecznych podań piłki oraz czasy biegu i pierwiastki kwadratowe tych liczb

SD	$\sqrt{s}$	$(\sqrt{s})^2$
-2	8.55	73.1
-1	9.01	81.2
0	9.47	89.7
1	9.93	98.6
2	10.39	108.0

**Tab. 3.5** Wartości średniej  $\pm$  jedno lub dwa odchylenia standardowe (pierwiastkowe) i odpowiadające im wartości po retransformacji (podniesione do kwadratu)

### *Transformacja zmiennych o skali zamkniętej*

Zmiennymi o skali bezwzględnie zamkniętej są np. „prawdziwe” procenty (skala od 0 do 100) i współczynniki korelacji (skala od  $-1$  do 1). Skala może być również zamknięta względnie (warunkowo), gdy wartości zmiennej nie mogą przybierać wartości poza określonym zakresem, jak np. maksymalna liczba punktów w kwestionariuszu psychometrycznym. We wszystkich tych sytuacjach wartości zmiennej będą miały rozkład normalny tylko wówczas, gdy wartość oczekiwana średniej jest w połowie skali, a więc np. 50%,  $r = 0$ , 23 punkty przy 45-punktowej skali itp. Gdy wartość oczekiwana jest inna, rozkład będzie tym bardziej prawo- lub lewoskośny, im bliżej lewego bądź prawego końca skali będzie leżała średnia.

Wartości procentowe obliczane są dla zmiennej dyskretnej, która ma inne właściwości niż zmienna ciągła. Na ogół, jeżeli obserwowane wartości mieszczą się w przedziale 30 – 70%, można nie dokonywać transformacji. Jeżeli jednak wartości wykraczają poza ten zakres, stosuje się najczęściej jedną z dwu transformacji – kątową lub logitową.

Przekształcenie kątowe ma postać 3-2, natomiast logitowe (dziesiętne) – 3-3.

$$k = \arcsin \sqrt{\%} \quad [3-2],$$

$$v = \log [\% / (100 - \%)] \quad [3-3]$$

Tablice tych wartości podane są na końcu książki (tablice 9 i 10).

Podany przykład dotyczy średniej ocen 12 studentów, obliczonej jako stosunek sumy wszystkich ocen z okresu studiów do maksymalnej możliwej sumy ocen (tab. 3.6); w lewej kolumnie podano obserwowane wartości procentowe, w środkowej – po przekształceniu kątowym, w prawej – po przekształceniu logitowym. W tabelce obok podane są średnie i odchylenia standardowe dla tych wartości oraz przedziały  $\bar{x} \pm 2 \text{SD}$ , a w dolnej części tej tabeli procenty obliczone zwrótnie (retransformowane) z odpowiednich wartości  $k$  i  $v$ . Jak widać, górna granica przedziału ufności obliczonego dla surowych wartości przekracza 100%, a więc przybiera nierealną wartość (nie może być wartości przekraczającej 100%, tak jak nie może być np. ujemnego czasu biegu).

%	k	v		$\bar{x}$	SD	$\bar{x} \pm 2\text{SD}$
57.0	47.0	0.12	%	76.9	13.5	49.9 – 103.9
60.2	50.9	0.18	k	62.21	10.14	41.9 – 82.5
63.3	52.7	0.24	v	0.61	0.41	-0.21 – 1.43
65.4	54.0	0.28	% (k)	78.3	–	44.6 – 98.3
70.3	57.0	0.37	% (v)	80.3	–	38.1 – 96.4
76.0	60.7	0.50				
79.5	63.1	0.59				
85.5	67.6	0.77				
86.4	68.4	0.80				
90.4	72.0	0.97				
92.7	74.3	1.10				
96.2	78.8	1.40				

**Tab. 3.6** Zaobserwowane procentowe wartości średnich ocen – surowe (%), po przekształceniu kątowym (k) i logitowym (v).

Z prawej – średnie, odchylenia i przedziały  $\bar{x} \pm 2 \text{SD}$ , oraz wartości retransformowane.

Przekształcenie współczynnika korelacji na zmienną o rozkładzie normalnym omówiono na s. 45 (wzór 5-2).



### Zmienna unormowana

Wartości unormowane ( $z_i$ ) oblicza się, dzieląc różnicę między danymi wartościami a średnią przez odchylenie standardowe:

$$z_i = \frac{x_i - \bar{x}}{SD} \quad [3-4]$$

W przykładzie poniżej (tab. 3.7) średnia wynosi 9.96, a odchylenie standardowe 0.52. Wartości unormowane wynoszą zatem:

$$z_1 = (10.3 - 9.96)/0.52 = 0.66, \quad z_3 = (9.1 - 9.96)/0.52 = -1.65, \text{ itd.}$$

W wyniku takiego normowania na wartości własne (średnią i SD) otrzymuje się średnią wartość unormowaną równą zero, a odchylenie standardowe równe jedności. Normując w ten sposób różne cechy otrzyma się dla każdej takie same wartości unormowane:  $\bar{x} = 0$ ,  $SD = 1$ , niezależnie od skal, w jakich były wyrażone ich wartości pierwotne, nieunormowane. Normowanie jest zatem sprowadzeniem wartości dowolnej cechy do jednej, wspólnej skali, wyrażonej w jednostkach odchyłeń standardowych tych cech, czyli w jednostkach funkcji  $t$ . Zachowane są przy tym proporcje pomiędzy poszczególnymi wartościami indywidualnymi.

nr	$x_i$	$z_i$
1	10.3	0.66
2	10.9	1.81
3	9.1	-1.65
4	10.1	0.27
5	9.7	-0.49
6	9.9	-0.11
7	9.7	-0.49
$\bar{x}$	9.96	0.00
SD	0.52	1.00

**Tab. 3.7** Przykład danych ( $x_i$ ) służących do obliczenia wartości unormowanych ( $z_i$ ) za pomocą wzoru 3-3

Normowanie może być także przeprowadzane na wartości zewnętrzne – np. na średnią i SD z grupy odniesienia (referencyjnej, kontrolnej), z norm opisanych w literaturze, itp. Jeżeli jest kilka grup, a nie ma wśród nich referencyjnej, można wówczas normować na średnią danej cechy z całego materiału i *średnie ważone* odchylenie standardowe ze wszystkich grup. Sposób obliczania średniej ważonej wariancji ( $SD^2$ ) podano w przykładzie jednokierunkowej analizy wariancji (s. 37). Można też stosować postępowanie mieszane, tzn. dla jednej cechy przeprowadza się normowanie na wartości z piśmiennictwa, dla innej – na średnią ogólną i średnie SD itp. Oczywiście, wartości, których rozkłady wskazują na potrzebę normalizacji, należy najpierw poddać transformacji (np. logarytmicznej), a dopiero potem unormować. Odpowiedni przykład pokazano w tabeli 3.8.

Średnia unormowana wartość dla ferrytyny wynosi zero, a  $SD = 1$ , ponieważ normowanie było przeprowadzone względem własnej średniej i odchylenia standardowego, nato-

miast dla pozostałych zmiennych, które normowano na wartości zewnętrzne, wartości te są inne, ale odchylenia standardowe (unormowane) są zbliżone do jedności. Oparcie normowania na średnich i odchyleniach standardowych sprawia, że unormowane zmienne mają właściwości *addytywne*, tzn. sumy (bądź średnie) różnych zmiennych unormowanych można traktować jako nową zmienną.

nr	Wartości pierwotne			Wartości unormowane		
	Ft	Fe-s	Fe-d	Ft	Fe-s	Fe-d
1	1.312	61	11.0	-1.35	-1.93	-2.45
2	1.812	83	14.1	1.31	-0.47	-0.55
3	1.661	98	15.3	0.51	0.53	0.18
4	1.485	70	12.4	-0.43	-1.33	-1.60
5	1.562	82	15.0	-0.02	-0.53	0.00
$\bar{x}$	<b>1.566</b>	78.8	13.56	0.00	-0.75	-0.88
SD	<b>0.188</b>	14.1	<b>1.63</b>	1.00	0.93	1.12
$\bar{x}$	-	<b>90</b>	<b>15</b>	← Normujące wartości zewnętrzne		
SD	-	<b>15</b>	-			

**Tab. 3.8** Przykłady normowania danych na średnie i odchylenia standardowe (wyfłuszczone) własne (wewnętrzne) lub zewnętrzne

Ft – stężenie ferrytyny w osoczu (log ng/ml), normowane na własną średnią i SD; Fe-s – stężenie żelaza w osoczu ( $\mu\text{g}/100\text{ ml}$ ), normowane na normy kliniczne; Fe-d – dobowe spożycie żelaza (mg), normowane na wartość zalecaną i odchylenie standardowe z próby

Wartości unormowane są wykorzystywane w analizie wieloczechowej, np. w analizie podobieństwa, analizie wariancji. Zastosowania zmiennych unormowanych będą omówione w rozdziale 6.

#### 4. JAK ZAPLANOWAĆ DOŚWIADCZENIE I OCENIĆ WYNIKI

Doświadczenia bądź obserwacje przeprowadza się zazwyczaj w celu stwierdzenia różnic między badanymi grupami, np. pod wpływem jakiegoś czynnika. Ocena wyników badań często jest utrudniona bądź wręcz niemożliwa, jeżeli badania zostały źle zaplanowane.

##### Porównanie dwu grup

Na początek rozpatrzmy prosty przykład: dwie grupy piłkarzy poddano testowi na prędkość biegu i porównano wyniki, a w drugim doświadczeniu badano prędkość biegu po tygodniowym treningu i porównano ją z prędkością uzyskaną przed rozpoczęciem treningu przez tych samych zawodników.

*Test t Studenta dla grup niezależnych (danych niezależnych, niesparowanych)*

W tabeli 4.1 podano dystans w metrach przebiegnięty w określonym czasie przez piłkarzy z grup A ( $n = 7$ ) i B ( $n = 8$ ); u dołu tabeli podane są sumy wartości oraz średnie i odchylenia standardowe. Różnica między średnimi ( $\Delta$ ) wynosi 61.56, a błąd standardowy tej różnicy ( $s_{\Delta}$ ) oblicza się z odchyłeń standardowych i liczebności (wzór 4-1):

	A	B
	123	255
	255	351
	240	292
	198	240
	198	225
	171	270
	240	270
		218
$\Sigma$	1425	2121
$\bar{x}$	203.57	265.13
SD	46.39	42.61

$$s_{\Delta} = \sqrt{S \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}$$

4-1

gdzie

$$S = [46.39^2 \cdot (7-1) + 42.61^2 \cdot (8-1)] / (7+8-2) = 25$$

621.5/13 = 1970.9 · (1/7 + 1/8) = 527.9 (jest to średnia wariancja), a pierwiastek tej liczby (błąd standardowy,  $s_{\Delta}$ ) wynosi 23.0.

**Tab. 4.1** Wyniki testu biegowego w dwu grupach piłkarzy

Wartość funkcji  $t$  Studenta dla różnicy między średnimi określona jest wzorem 4-2.

$$t = \frac{\Delta}{s_{\Delta}}$$

[4-2]

W powyższym przykładzie wynosi ona  $61.56/23.0 = 2.68$ . Ponieważ liczba stopni swobody ( $df$ ) dla  $s_{\Delta}$  wynosi  $(7-1) + (8-1) = 13$ , a wartość  $t$  dla  $df = 13$  i  $P = 0.05$  wynosi 2.16 (zob. tab. 3, s. 88), uznajemy wynik za znamiennej (nieprzypadkowy) na poziomie  $P < 0.05$ . Innymi słowy, średni wynik testu biegowego w grupie B jest znamiennej wyższy niż w grupie A.

Należy zaznaczyć, że test  $t$  może być użyty **wyłącznie** wówczas, gdy porównuje się dwie grupy. Wprawdzie przy np. trzech grupach porównuje się średnie parami (A-B, A-C, B-C), ale każda średnia występuje dwukrotnie, a więc porównania nie są **niezależne**. Przy większej liczbie grup należy zatem zastosować analizę wariancji (patrz niżej), a następnie jakiś test 'wielokrotnego rozstępu', np. test Scheffé'go.

Powyższe obliczenia będą poprawne, jeżeli odchylenia standardowe obu grup są podobne. Przy znacznej (np. ponaddwukrotnej) różnicy między odchyleniami standardowymi bezpieczniej jest użyć testu nieparametrycznego (zob. s. 71) lub zbadać, czy wariancje różnią się znamiennej stosując np. test Bartletta (zob. s. 29). Bardziej skomplikowane metody parametryczne nie będą tu omawiane.

#### *Test t Studenta dla danych zależnych (sparowanych)*

Często przeprowadza się pomiary u tych samych osób w dwu różnych sytuacjach, np. przed i po wysiłku. Są to tzw. próby zależne. Najprostszy sposób polega na obliczeniu indywidualnych różnic między drugim i pierwszym pomiarem, a następnie średniej i odchylenia standardowego tych różnic, a z tego ostatniego – błędu standardowego ( $SE = SD/\sqrt{n}$ ). Wartość testu  $t$  jest równa ilorazowi średniej różnicy i SE.

Test biegowy wykonany przez piłkarzy z grupy B z poprzedniego przykładu (tab. 4.1, kolumna B) został powtórzony po tygodniowym treningu (kolumna II w tab. 4.2). W kolumnie  $\Delta$  tej tabeli podane są różnice między wynikami testu II i I. Obliczamy średnią indywidualną różnicę i odchylenie standardowe – wynoszą one  $19.5 \pm 20.3$ . Błąd standardowy średniej różnicy wynosi:

$$s_{\Delta} = 20.3/\sqrt{8} = 7.18, \text{ zatem } t = 19.5/7.18 = 2.72.$$

I	II	$\Delta$
255	302	47
351	344	-7
292	315	23
240	251	11
225	240	15
270	318	48
270	292	22
218	215	-3
n = 8	$\bar{x}$	19.5
	SD	20.3

Wartość tablicowa dla  $df = 7$  i  $P = 0.05$  wynosi 2.37, uznajemy zatem, że zastosowany trening wywołał znamiennej wzrost sprawności biegowej w tej grupie piłkarzy.

**Tab. 4.2** Wyniki testu biegowego przeprowadzonego dwukrotnie u tych samych osób i różnice między powtórzeniami ( $\Delta$ )

## Analiza wariancji (ANOVA)

Analiza wariancji polega na określeniu zróżnicowania średnich wartości danej cechy w kilku grupach. Zróżnicowanie to wyraża się wartością funkcji  $F$  Snedecora, która jest stosunkiem wariancji międzygrupowej (między średnimi) do wariancji wewnątrzgrupowej („resztowej”). Znamienne wartości funkcji  $F$  oznaczają, że badane średnie nie pochodzą z jednorodnej populacji (a więc są znamienne zróżnicowane), ale nie wskazuje, między którymi konkretnymi średnimi różnica jest znamienne. Może być tak, że w zespole kilku średnich tylko jedna będzie się znamienne różnić od pozostałych, albo że różnić się będą tylko najmniejsza i największa wartość, itp. Jeżeli natomiast wynik analizy wariancji jest nieznamienne, to znaczy, że żadne dwie średnie nie różnią się między sobą. Jeżeli zatem wynik analizy wariancji jest znamienne, należy przeprowadzić dalsze testy (tzw. testy *post-hoc*) dla stwierdzenia, między którymi średnimi różnice są znamienne.

Aby można było przeprowadzić analizę wariancji, powinny być spełnione pewne warunki: rozkłady danej cechy w badanych grupach nie mogą zbytnio odbiegać od rozkładu normalnego, a wariancje w poszczególnych grupach powinny być podobne. Na ogół wystarcza spełnienie tego ostatniego warunku, który sprawdza się za pomocą *testu Bartletta* (patrz dalej).

Analiza wariancji jest wykorzystywana w bardzo wielu układach doświadczalnych o różnym stopniu komplikacji, od najprostszej jednokierunkowej, polegającej na porównaniu średnich z kilku niezależnych grup, po bardzo złożone wielowymiarowe schematy. Poniżej zostanie omówiona jednokierunkowa analiza oraz niektóre formy dwukierunkowej analizy wariancji.

### *Jednokierunkowa analiza wariancji*

W tabelce poniżej podano średnie i odchylenia standardowe mocy maksymalnej zmierzonej w teście Wingate w trzech grupach zawodników; pytanie brzmi, czy średnie wartości mocy maksymalnej w tych trzech grupach różnią się znamienne między sobą.

	n	$\bar{x}$	SD
A	11	9.03	0.409
B	8	10.77	0.646
C	5	9.88	0.317

**Tab. 4.3** Średnie i odchylenia standardowe mocy maksymalnej w trzech grupach zawodników

Zanim przystąpi się do analizy, należy zwrócić uwagę na to, że odchylenie standardowe w grupie B jest dwukrotnie wyższe niż w grupie C (a to znaczy, że wariancja w grupie B jest 4-krotnie większa niż w grupie C, bo wariancja jest kwadratem odchylenia standardowego). Należy zatem najpierw sprawdzić, czy warunek równości wariancji w badanych

grupach można uznać za spełniony. W tym celu, korzystając z danych zawartych w tabeli 4.3, należy obliczyć średnią ważoną wariancję  $\bar{s}^2 = \frac{4.99589}{21} = 0.2379$  i jej logarytm (-0.6236),

a potem wartość  $M$ :

$$M = 2.3026 \cdot (\log \bar{s}^2 \cdot \Sigma df - \Sigma df \cdot \log s^2)$$

gdzie 2.3026 jest stałą, równą  $\ln(10)$ . Wartość  $M$  odpowiada funkcji chi-kwadrat.

	df	s	s <sup>2</sup>	df · s <sup>2</sup>	df · log s <sup>2</sup>	n	$\bar{x}$	n · $\bar{x}$	n · $\bar{x}^2$
A	10	0.409	0.16728	1.67281	-7.7655	11	9.03	99.33	896.95
B	7	0.646	0.41732	2.92121	-2.6567	8	10.77	86.16	927.943
C	4	0.317	0.10049	0.40196	-3.9915	5	9.88	49.4	488.072
Σ	21	1.372	0.68509	4.99598	-14.414	24		234.89	2312.97

**Tab. 4.4** Dane dla trzech grup potrzebne do testu Bartletta i przeprowadzenia analizy wariancji

W powyższym przykładzie  $M = 2.3026 \cdot (-0.6236 \cdot 21 + 14.414) = 3.04$ ; wartość ta jest mniejsza od tablicowej wartości  $\chi^2$  dla  $df = k-1$  ( $k$  – liczba grup) wynoszącej 5.99, a więc warunek równości wariancji w badanych grupach można uznać za spełniony dla potrzeb analizy wariancji.

Uproszczony rachunek analizy wariancji (ze średnich i SD) przedstawia się następująco:

Obliczona powyżej wariancja resztowa wynosi 0.2379, suma kwadratów “międzygrupowa” wynosi  $2312.97 - 234.89^2/24 = 14.08$  dla  $df = k-1$ , zatem wariancja międzygrupowa wynosi  $14.08/2 = 7.04$ .

Stąd, wartość funkcji  $F = 7.04/0.2379 = 29.6$ , co znacznie przekracza wartość tablicową dla 2 i 21 stopni swobody (5.80, dla  $P = 0.01$ ), a więc zróżnicowanie między średnimi jest wysoce znamienne. Dane te zestawiono w tabeli 4.5.

Zmienność	S	df	s <sup>2</sup>	F
Grupy	14.08	2	7.04	29.6
Reszta	4.996	21	0.2379	

**Tab. 4.5** Tablica analizy wariancji dla danych z tabeli 4.4 i zawartych w tekście

Należy teraz zbadać, między którymi średnimi (A-B, A-C, B-C) różnice są znamienne. Oceny zmienności różnic dokona się za pomocą testu Scheffé’go, polegającego na uszeregowaniu wszystkich średnich od najmniejszej do największej, a następnie na porównaniu wszystkich różnic z odpowiadającymi im *najmniejszymi istotnymi różnicami* (NIR),

przy czym NIR jest określona wzorem 4-3.

(objaśnienia symboli w przykładzie poniżej).

$$\boxed{\text{NIR} = s_{\Delta} \cdot \sqrt{\mathbf{h} \cdot \mathbf{F}_{h,r}}} \quad [4-3]$$

Kolejność średnich w powyższym przykładzie jest następująca:

9.03 (A), 9.88 (C) i 10.77 (B), a wariancja resztowa wynosi 0.2379 przy  $df = 21$ .

Kolejność obliczeń jest następująca:

różnica między średnimi B i A wynosi  $10.77 - 9.03 = 1.74$ , błąd tej różnicy ( $s_{\Delta}$ ) zgodnie z wzorem [4-1] wynosi

$$\sqrt{0.238 \left( \frac{1}{11} + \frac{1}{8} \right)} = 0.227$$

zaś  $NIR = 0.227 \cdot \sqrt{2} \cdot 3.47 = 0.60$ .

We wzorze tym '2' (pod pierwiastkiem) jest różnicą między kolejnymi numerami porównywanych średnich (w uporządkowanym szeregu średnia B ma numer 3, a średnia A – 1), a 3.47 jest wartością funkcji  $F$  Snedecora dla 2 i 21 stopni swobody przy  $P=0.05$  (por. tablica 4). Ponieważ różnica między średnimi B i A jest większa niż NIR, uznajemy ją za znamiennej na poziomie  $P<0.05$ .

Łatwo zauważyć, że porównując dwie kolejne średnie wartość  $h$  we wzorze na NIR wynosi 1, więc  $NIR = s_{\Delta} \cdot \sqrt{F_{1,r}}$ , zaś pierwiastek z  $F$  dla  $df = 1$ ,  $r$  jest równy  $t$  dla  $df = r$ . W uporządkowanym szeregu średnich, każde dwie średnie sąsiadujące ze sobą (a więc A/C lub C/B w powyższym przykładzie) można porównać za pomocą testu  $t$ , natomiast do porównania średnich niesąsiadujących (A/B) należy użyć testu Scheffé'go.

Aby lepiej zilustrować jednokierunkową analizę wariancji, rozpatrzmy przykład 4 grup, w których oznaczono poziom łęku, biorąc za punkt wyjścia wynik analizy wariancji bez poprzedzających obliczeń.

Grupa	n	$\bar{x}$	SD
A	11	43.8	5.83
B	19	51.5	5.06
C	22	46.8	5.70
D	20	49.0	4.39

**Tab. 4.6** Średnie i odchylenia standardowe wartości poziomu łęku w 4 grupach zawodników.  
Wariancja resztowa ( $s^2$ ) = 27.2 przy  $df = 68$ ,  $F_{3,68} = 5.83$  ( $P<0.001$ )

Ponieważ zróżnicowanie między średnimi jest wysoce znamienne, szeregujemy średnie od najmniejszej do największej (A – C – D – B) i obliczamy wartości NIR dla kolejnych różnic korzystając z wzoru [4-3]; wartość  $5.22 = \sqrt{27.2}$ .

$$NIR_{AB} = 5.22 \sqrt{3 \cdot 2.75 \cdot \left( \frac{1}{11} + \frac{1}{19} \right)} = 5.68 \quad NIR_{AC} = 5.22 \sqrt{1 \cdot 3.99 \cdot \left( \frac{1}{11} + \frac{1}{22} \right)} = 3.85$$

$$NIR_{AD} = 5.22 \sqrt{2 \cdot 3.14 \cdot \left( \frac{1}{11} + \frac{1}{20} \right)} = 4.91 \quad \text{itd.}$$

Różnice między średnimi wynoszą:

$$B - A = 7.7, \quad D - A = 5.2, \quad C - A = 3.0,$$

a więc ta ostatnia, jako mniejsza od  $NIR_{AC}$ , jest nieznamiennej. Po wykonaniu wszystkich rachunków (6 różnic) okazuje się, że znamienne są różnice: B – A, C – A i B – C. Wynik ten zapisuje się w skróconej formie jak poniżej. Grupy, których średnie nie różnią się między sobą, podkreśla się.

1	2	3	4
A	C	D	B

Pokazane tu rachunki wykonane na średnich i odchyleniach standardowych są oczywiście uproszczone. Ponieważ błędy zaokrągleń mogą mieć znaczny wpływ na końcowy wynik analizy, wartości odchyżeń standardowych powinny być obliczone z dobrą dokładnością (4 miejsca znaczące). Dokładne obliczenia oparte na indywidualnych danych przedstawiono poniżej dla układu *cross-over*.

#### *Układ cross-over*

W przykładzie dotyczącym obliczeń dla dwu grup wyników w układzie zależnym (s. 34), kolejność „zabiegów” (spoczynek, wysiłek) jest narzucona przez sytuację, bowiem pomiar spoczynkowy po uprzednim wykonaniu wysiłku mógłby być obciążony dużym błędem. Rozpatrzmy jednak inną sytuację, np. wpływ diety „kontrolnej” i „doświadczalnej” na wydolność. Mimo że kolejność diet może być dowolna, zwykle stosuje się u wszystkich badanych najpierw kontrolną, a potem doświadczalną. Jest to błąd, gdyż nie ma wówczas możliwości wykrycia lub uniknięcia tzw. działania następczego. Właściwym sposobem jest tu zastosowanie układu *cross-over*. Polega on na tym, że badaną grupę dzieli się najpierw losowo na pół, a następnie jednej podgrupie podaje się najpierw dietę A, a potem B, a drugiej – najpierw dietę B, potem A.

Przykład takiego układu doświadczalnego: w tabeli 4.7 przedstawiono czas utrzymania mocy maksymalnej w teście Wingate po zastosowaniu dwu różnych form treningu – A i B. Każdy trening prowadzony był przez 2 tygodnie z tym, że w grupie I zastosowano najpierw trening A, a potem B, a w grupie II najpierw trening B, potem A. Schemat ten pozwala ocenić wpływ następczości zabiegów, a więc wpływ poprzedniego zabiegu na następny.

I	A	B	$\Sigma$
1	4.4	3.8	8.2
2	4.5	4.2	8.7
3	3.6	2.9	6.5
4	3.6	3.9	7.5
5	4.0	3.4	7.4
$\Sigma$	20.1	18.2	38.3
II	B	A	$\Sigma$
6	2.7	2.6	5.3
7	3.3	3.7	7.0
8	3.4	3.7	7.1
9	3.1	3.8	6.9
10	2.2	3.1	5.3
$\Sigma$	14.7	16.9	31.6

**Tab. 4.7** Wartości czasu utrzymania mocy maksymalnej w teście Wingate po stosowaniu dwu form treningu – A i B



Aby wykonać analizę wariancji, należy obliczyć następujące wielkości:

- całkowitą sumę kwadratów ( $\Sigma_C$ ),
- sumę kwadratów dla osób ( $\Sigma_O/2$ ; bo są 2 wyniki dla każdej osoby),
- sumę kwadratów dla treningów ( $\Sigma_T/10$ ; bo jest 10 wyników z każdego treningu),
- sumę kwadratów dla kolejności zabiegów (następczości)  $\Sigma_N/10$  (jak wyżej),
- poprawkę (P), którą odejmuje się od każdej z powyższych wielkości. Otrzymuje się wówczas tzw. zredukowane sumy kwadratów ( $S_{xx}$ ):  $C = \Sigma_C - P$ ,  
 $O = \Sigma_O - P$ ,  $T = \Sigma_T - P$ ,  $N = \Sigma_N - P$ . Obliczenia wykonuje się następująco:

$P = 69.9^2/20 = 244.300$	
$\Sigma_C = 4.4^2 + 3.8^2 + 4.5^2 + \dots + 3.1^2 = 251.170$	$C = 6.870$
$\Sigma_O = (8.2^2 + 8.7^2 + \dots + 5.3^2)/2 = 249.695$	$O = 5.395$
$\Sigma_T = [(20.1+16.9)^2 + (18.2+14.7)^2]/10 = 245.141$	$T = 0.841$
$\Sigma_N = [(20.1+14.7)^2 + (18.2+16.9)^2]/10 = 244.305$	$N = 0.004$
$R$ ('resztowa') – z różnicy: $R = C - O - T - N$	

W tablicy analizy wariancji umieszcza się następujące dane:  $S_{xx}$ , liczbę stopni swobody (df) równą liczbie elementów minus jeden, wariancję ( $S_{xx}/df$ ), wreszcie funkcję  $F$  Snedecora, z podzielenia danej wariancji przez wariancję resztową; np., dla zmienności między rodzajami treningów  $F = 0.841/0.079$ .

	$S_{xx}$	df	$s^2$	F
C (całkowita)	6.870	19	-	-
O (osoby)	5.395	9	0.599	7.61
T (treningi)	0.841	1	0.841	10.67
N (następczość)	0.004	1	0.004	<1
reszta	0.630	8	0.079	-

**Tab. 4.8** Tablica analizy wariancji dla danych z tabeli 4.7

W powyższym przykładzie wystąpiła znamienna różnica między rodzajami treningu – trening A był skuteczniejszy niż trening B dla dłuższego utrzymywania mocy maksymalnej w wysiłku maksymalnym ( $F = 10.67$ ,  $P < 0.05$ , dla  $df = 1$  i  $8$ ). Przy tym nie wystąpiło działanie następcze ( $F < 1$ ), czyli kolejność stosowanych treningów nie miała znaczenia. Gdyby jednak było inaczej, to zastosowanie innego, prostszego układu doświadczalnego nie pozwoliłoby tego efektu wykryć.

Układ *cross-over* może być stosowany do więcej niż dwóch wariantów doświadczalnych; trzeba wówczas podzielić wszystkich badanych na tyle grup, ile jest wariantów i odpowiednio zmieniać kolejność zabiegów. W takim wypadku właściwszy byłby układ *kwadratów łacińskich*, który nie będzie tu jednak omawiany.

### Układ krzyżowy

Układ taki polega na tym, że z jednej strony mamy podział na pewne kategorie (A), np. płeć, różne grupy społeczne lub dyscypliny sportowe itp., a z drugiej – inne kategorie (B), np. rodzaj spożywanego diety, występowanie określonych schorzeń itp.

W poniższym przykładzie jednym kierunkiem podziału jest pochodzenie klubu rugby, drugim – formacja rugbistów, przy czym obie formacje w danym klubie są jednakowo liczne. Taki układ doświadczalny jak w tabeli 4.11 nazywamy krzyżowym, a stosuje się do niego **dwukierunkową analizę wariancji** (bo mamy dwa kierunki zmienności – kluby i formacje). Gdyby liczebności w każdej klatce tabeli (‘obiekcie’) były jednakowe i nie brakowałoby żadnych danych, byłaby to *tabela ortogonalna*. Wprawdzie całkowite liczebności poszczególnych wierszy są różne, ale liczebności obiektów w danym wierszu są jednakowe. Taka tabela może być zatem uważana za ortogonalną z punktu widzenia obliczeń.

Formacja	Obrona		Atak	
	n	średnia $\pm$ SD	n	średnia $\pm$ SD
Pd. Afryka	14	55.8 $\pm$ 4.1	14	52.0 $\pm$ 4.8
USA	17	59.5 $\pm$ 3.0	17	54.1 $\pm$ 2.9
Walia	19	59.1 $\pm$ 2.8	19	54.3 $\pm$ 3.1

**Tab. 4.9** Maksymalne pochłanianie tlenu ( $\dot{V}O_2\max$ , ml/min/kg) zmierzone u piłkarzy obrony i ataku z trzech klubów

Ponieważ obliczenia na indywidualnych wynikach zostały pokazane na przykładzie układów *cross-over* i hierarchiczno-krzyżowego, ograniczymy się tutaj do pokazania jedynie końcowej tablicy analizy wariancji dla danych z tabeli 4.11. Ponieważ dane (średnie i SD) w każdym obiekcie pochodzą od różnych osób, tablica analizy wariancji będzie zawierać jedynie zmienność między wierszami (klubami), między kolumnami (formacjami) oraz interakcję wiersze  $\times$  kolumny.

	$S_{xx}$	df	$s^2$	F
Całkowita	1842.13	99		
Obiekty	731.41	5	146.28	
Kluby	557.88	1	557.88	47.21**
Formacje	163.59	2	81.80	13.85**
Kluby $\times$ Form.	9.94	2	4.97	<1
Reszta	1110.71	94	11.82	

**Tab. 4.10** Tablica analizy wariancji dla danych z tabeli 4.11  
\*\*  $P < 0.01$

Analizę wyników zaczynamy, jak zawsze, od interakcji. Ponieważ jest ona nieznamienista, przyjmujemy, że formacje różniły się między sobą znamienne (F = 13.85) i podobnie we wszystkich klubach. Znamienista interakcja świadczyłaby o tym, że różnice między formacjami pod względem  $\dot{V}O_2\max$  są różne w różnych klubach (np. w jednym klubie

zawodnicy ataku mieliby niższe  $\dot{V}O_2\text{max}$  niż zawodnicy obrony, a w innym klubie – odwrotnie). Otrzymany wynik mówi, że we wszystkich badanych klubach obrona ma wyższe  $\dot{V}O_2\text{max}$  niż atak. Natomiast znamienne zróżnicowanie między klubami ( $F = 47.21$ ) nie świadczy o tym, że każdy klub jest inny: z tablicy 4.10 wynika, że kluby USA i Walii są niemal identyczne, a klub z Południowej Afryki ma wyniki proporcjonalnie niższe w obu formacjach.

	$S_{xx}$	df	$s^2$	F
Całkowita	1059.59	71		
Obiekty	466.94	3	155.65	
Kluby	465.13	1	465.13	53.37
Formacje	0.19	1	0.19	<1
Kluby×Form.	1.63	1	1.63	<1
Reszta	592.66	68	8.72	

**Tab. 4.11** Tablica analizy wariancji dla klubów USA i Walii (dane z tabeli 4.11)

Brak znamienych różnic między klubami USA i Walii można wykazać powtarzając analizę tylko dla tych klubów (tab. 4.13); jak widać, zróżnicowanie między klubami jest nieznamienne ( $F < 1$ ).

Nieortogonalność tabeli wyników, a więc niejednakowe liczebności obiektów, bardzo komplikuje obliczenia, dlatego planując doświadczenie dwu- (lub więcej) kierunkowe, należy uwzględnić ortogonalność. Jeżeli jednak układ doświadczalny zawiera nie więcej niż dwa wiersze (liczba kolumn może być dowolna), rachunek jest stosunkowo mało skomplikowany, jak pokazano na poniższym przykładzie.

Grupy	Nietrenujące		Zawodniczki	
	n	średnia $\pm$ SD	n	średnia $\pm$ SD
Zasoby żelaza				
Zasoby żelaza w normie	20	12.5 $\pm$ 2.3	45	13.4 $\pm$ 4.3
Niedobór żelaza	22	9.6 $\pm$ 2.4	15	14.1 $\pm$ 4.2
w, $\Delta$	10.48	2.9	11.25	0.7

**Tab. 4.12** Dobowe spożycie żelaza niehemowego w grupach kobiet nietrenujących i zawodniczek, wykazujących prawidłowe zasoby lub niedobory żelaza w ustroju

Analizę zaczyna się od obliczenia interakcji. W tym celu należy wyznaczyć dla każdej kolumny wagę ( $w$ ) i różnicę między średnimi w obu wierszach ( $\Delta$ ), a następnie wielkości  $w \cdot \Delta$  i  $w \cdot \Delta^2$ . Dwie ostatnie wartości oraz wagi sumuje się dla każdego wiersza przez wszystkie kolumny. Wartości te dla pierwszej kolumny ('nietrenujące') wynoszą:  $\Delta = 12.5 - 9.6 = 2.9$ ;  $w = 20 \cdot 22 / (20 + 22) = 10.48$ ;  $w \cdot \Delta = 30.39$ ;  $w \cdot \Delta^2 = 88.14$ , sumy dla obu kolumn zaś wynoszą:  $\sum w = 21.73$ ;  $\sum w \cdot \Delta = 22.52$ ;  $\sum w \cdot \Delta^2 = 93.65$ . Interakcja jest równa  $\sum w \cdot \Delta^2 - \sum W \cdot \Delta = 93.65 - 22.52 = 71.13$  (na jednym stopniu swobody), suma 'resztowa' zaś jest

równa sumie  $S_{xx}$  z poszczególnych klatek, czyli  $\sum SD^2 \cdot (n-1)$ , gdzie SD, n są to odchylenia standardowe i liczebności dla poszczególnych klatek.

Suma resztowa wynosi zatem:  $19 \cdot 2.3^2 + 44 \cdot 4.3^2 + 21 \cdot 2.4^2 + 14 \cdot 4.2^2 = 1281.99$ , wariancja resztowa zaś wynosi  $1281.99/98 = 13.08$ . Dzieląc interakcję przez wariancję resztową otrzymamy:  $F = 71.13/13.08 = 5.44$  (df = 1 i 98;  $P < 0.05$ ). Ponieważ interakcja jest znamienna, nie oblicza się zróżnicowania między wierszami oraz kolumnami, lecz ocenia różnice między średnimi w wierszach dla każdej kolumny oddzielnie za pomocą testu  $t$ , jak na stronie 26.

#### Układ bloków losowych

Niekiedy nie ma możliwości zastosowania określonych zabiegów i przeprowadzenia pomiarów u tych samych osób, a podejrzewamy, że byłby to właściwszy model doświadczalny niż grupy niezależne, np. z powodu znacznej niejednorodności badanej grupy i związanej z tym dużej zmienności, co mogłoby się odbić na wynikach. W takich wypadkach stosuje się tzw. układ *bloków losowych*. Na przykład:

Blok	Trening A		Trening B		Suma mocy
	m.c.	moc	m.c.	moc	
1	73	10.2	73	9.3	19.5
2	75	10.5	76	9.1	19.6
3	77	10.0	77	8.5	18.5
4	78	10.2	78	10.6	20.8
5	78	9.0	78	8.9	17.9
6	80	9.8	80	8.9	18.7
7	80	9.4	80	9.0	18.4
8	82	8.9	83	8.6	17.5
9	84	9.3	84	9.5	18.8
$\bar{x}$		9.70		9.16	

**Tab. 4.13** Masa ciała (m.c.) i moc względna (W/kg) w krótkim teście na cykloergometrze po treningu A lub B. Przed treningiem zawodnicy zostali podzieleni na bloki pod względem masy ciała.

Badano wpływ dwóch rodzajów treningu siłowego (A i B) przeprowadzonego na zgrupowaniu na moc osiąganą w krótkim teście na cykloergometrze. Pomiaru mocy odnoszono do masy ciała badanych. Ponieważ nie było możliwości przeprowadzenia testu na zgrupowaniu przed treningiem, aby bezpośrednio zmierzyć zmiany mocy wywołane treningiem, wybrano 9 par zawodników tak, by w każdej parze masy ciała były jak najbardziej podobne. Następnie z każdej pary wybrano **losowo** jednego zawodnika do grupy A, a pozostałego przydzielono do grupy B. Po dwutygodniowym treningu i powrocie do Warszawy wszystkich zawodników poddano testowi. Masy ciała zawodników i wyniki testu przedstawiono w tabeli 4.13.

	$S_{xx}$	df	$s^2$	F		$S_{xx}$	df	$s^2$	F
Całkowita	7.08	17				7.08	17		
Bloki	3.94	8	0.493	2.19		–	–	–	–
Grupy	1.34	1	1.340	5.96*		1.34	1	1.340	3.73
Reszta	1.80	8	0.225			5.74	16	0.359	

**Tab. 4.14** Tablice analizy wariancji: dla układu bloków losowych (z lewej) i bez uwzględnienia podziału na bloki losowe (z prawej)

Po wydzieleniu w analizie wariancji składowej ‘bloki’ wariancja resztowa jest na tyle mała (0.225), że różnica między grupami (9.70 i 9.16) jest znamienna ( $F = 5.96$ ,  $P < 0.05$ ), mimo że wariancja resztowa ma tylko 8 stopni swobody. Gdyby nie utworzono bloków losowych tylko potraktowano obie grupy zawodników jako niezależne, zmienność międzyosobnicza weszłaby do resztowej, której wariancja byłaby zbyt duża, by różnica między średnimi grup była znamienna ( $F = 3.73$ ,  $P > 0.05$ ). Wprawdzie wariancja ‘bloki’ nie jest znamienna względem resztowej ( $F = 2.19$ ,  $P > 0.05$ ), nie ma zatem podstaw do wydzielenia tej składowej, ale widoczna jest tendencja, która byłaby znamienna przy większej liczbie bloków.

Należy tu zauważyć, że szczególnym przypadkiem bloków losowych jest stosowanie różnych zabiegów na tych samych osobach – każda osoba stanowi wówczas blok losowy.

#### *Układ hierarchiczno-krzyżowy*

Doświadczenie przeprowadza się na różnych grupach, ale w każdej grupie są takie same zabiegi (lub np. okresy badań). W tabeli 4.15 przedstawiono wyniki pomiarów maksymalnego pochłaniania tlenu (ml/min/kg) w 3 grupach zawodników (A, B, C) w 2 okresach badań (I, II) – przed i po miesięcznym treningu.

Grupa	I	II	$\Sigma$
<b>A</b>	47.7	56.2	103.9
	55.6	62.9	118.5
	57.4	60.2	117.6
	52.1	58.2	110.3
	$\Sigma$	212.7	237.5
średnie	53.2	59.4	56.3
<b>B</b>	57.6	65.2	122.8
	52.7	60.1	112.8
	54.3	61.8	116.1
	53.0	63.8	116.8
	52.3	62.2	114.5
$\Sigma$	269.8	313.1	582.9
średnie	54.0	62.6	58.3
<b>C</b>	57.6	63.3	120.9
	70.6	75.0	145.6
	55.5	62.8	118.3
	58.3	63.7	122.0
	60.0	67.8	127.8
	56.8	63.4	120.2
	56.8	60.4	117.2
$\Sigma$	415.6	456.4	872.0
średnie	59.4	65.2	62.3
$\Sigma\Sigma$	898.2	1007.0	1905.2
średnie	56.14	62.94	

Suma kwadratów (SK) całkowita (C):

$$47.7^2 + 55.6^2 + \dots + 60.4^2 = 114\,428.5$$

SK dla osób (O):

$$(103.9^2 + 118.5^2 + \dots + 117.2^2) / 2 = 114\,024.9$$

SK dla „obiektów” (B):

$$(212.7^2 + 237.5^2) / 4 + (269.8^2 + 313.1^2) / 5 + (415.6^2 + 456.4^2) / 7 = 114\,016.2$$

SK dla grup (G):

$$450.2^2 / 8 + 582.9^2 / 10 + 872.0^2 / 14 = 113\,133.4$$

SK dla okresów (K):

$$(898.2^2 + 1007.0^2) / 16 = 113\,797.4$$

Od każdej z powyższych sum odejmujemy poprawkę

(P):

$$1905.2^2 / 32 = 113\,427.3$$

**Tab. 4.15** Pomiar maksymalnego pochłaniania tlenu (ml/min/kg) w 3 grupach zawodników (A, B, C) w 2 okresach badań: przed (I) i po (II) miesięcznym treningu

źródło zmienności	$S_{xx}$	df	$s^2$	$F_r$	$F_{og}$
Całkowita	1001.3	31			
Osoby	597.6	15			
Grupy	206.1	2	103.1		3.43
Osoby w Grupach	391.5	13	30.1	18.7*	
oBiekty	588.9	5			
[grupy]	[206.1]	[2]		(powtórzone)	
oKresy	370.1	1	370.1	230* 165*	
Grupy $\times$ oKresy	12.7	2	6.3	<1	
reszta (C-O-B)	20.9	13	1.61		
reszta (C-O-K)	33.6	15	2.24		

**Tab. 4.16** Analiza wariancji dla danych z tabeli 4.15

Tabela 4.16 zawiera pełną analizę wariancji dla omówionych danych. Zmienność resztowa jest różnicą całkowitej, międzysobniczej i „obiektywnej” (bez międzygrupowej, która już raz weszła do międzysobniczej). Zmienność resztowa w ostatnim wierszu (całkowita – osoby - okresy) odpowiada tej, która wynika z indywidualnych różnic między okresami ocenianymi testem  $t$  dla danych sparowanych.

Wartość  $F$  dla grup oraz interakcji grupy $\times$ okresy obliczana jest względem wariancji ‘międzyosobniczej w grupach’ (OG), natomiast pozostałe – względem wariancji resztowej.

Zróznicowanie między grupami jest znamienne na poziomie  $P < 0.10$ , co wskazuje na tendencję do różnic. Znamienność byłaby większa przy liczniejszych grupach. Sposób oszacowania liczebności próby niezbędnej do uzyskania znamiennej różnicy omówiono dalej (s. 48).

Efektywność schematu doświadczalnego (próby zależne bądź niezależne w badaniach typu przed/po, np. wysiłku) zależy od tego, czy wynik „po zabiegu” może być skorelowany z wynikiem „przed zabiegiem”, czy nie. Na przykład, powysiłkowa aktywność kinazy kreatynowej w surowicy jest tym większa, im większa jest wartość spoczynkowa. W takim wypadku należy użyć próby zależnej, np. bloków losowych, bowiem uzyskany efekt będzie wyraźniejszy niż dla prób niezależnych. W razie braku takiej korelacji między wynikami „przed” i „po”, bardziej skuteczny będzie układ prób niezależnych. Oczywiście, jeżeli po przeprowadzeniu doświadczenia np. na grupach niezależnych okazało się, że właściwszy byłby model zależny, znaczy to tylko, że zastosowano niewłaściwy schemat badań, co nie ma jednak wpływu na sposób obliczeń.

### Jak wstawić brakujące dane

Mając do czynienia nie z szeregiem, lecz z tablicą danych, np. w analizie wariancji, analizie wielocechowej itp., zdarza się niekiedy, że brak jednej lub kilku wartości. Jeżeli liczba wierszy w tablicy (np. liczba badanych osób) jest duża, można sobie pozwolić na usunięcie całego wiersza, w którym brakuje wartości, zwykle jednak szkoda traconych w ten sposób danych. Aby tego uniknąć, stosuje się wstawianie oszacowanych wartości w puste miejsca tablicy, jak pokazano w poniższym przykładzie.

nr	I	II	$\Sigma$
1	47.7	56.2	103.9
2	55.6		55.6
3	57.4	60.2	117.6
4	52.1	58.2	110.3
$\Sigma$	212.8	174.6	387.4

**Tab. 4.17** Dane z tabeli 4.15 (grupa A) z brakującym wynikiem

Brakujący wynik oblicza się z wzoru 4-4, gdzie  $W$  jest sumą wiersza z brakującym wynikiem,  $K$  – sumą kolumny z brakującym wynikiem,  $C$  – całkowitą sumą,  $w$  – liczbą wierszy,  $k$  – liczbą kolumn.

$$x = \frac{W \cdot w + K \cdot k - C}{(w-1) \cdot (k-1)} \quad 4-4$$

Brakująca wartość wynosi zatem  $(55.6 \cdot 4 + 174.6 \cdot 2 - 387.4)/(3 \cdot 1) = 61.4$ , co dość dobrze zgadza się z faktyczną wartością 62.9 (por. tab. 4.15).

Jeżeli w tablicy danych brak więcej niż jednej wartości, należy pamiętać, że w danym wierszu i kolumnie można wstawić nie więcej niż jedną wartość. Ponadto, od liczby stopni swobody dla błędu (wariancji resztowej) należy odjąć liczbę wstawionych wartości. Gdyby w przykładzie z tab. 4.15 rzeczywiście brakowało jednej wartości i zostałyby ona wstawiona, wówczas  $df$  dla wariancji resztowej wyniosłoby  $13 - 1 = 12$  (zob. tab. 4.16).

### **Wymagana liczebność grupy**

Niekiedy zdarza się, że obliczona różnica między średnimi nie osiąga znamienności, gdyż przyjęto zbyt małą liczebność grup i zachodzi pytanie, jaka powinna być liczebność, aby dana różnica była znamienna. Aby odpowiedzieć na to pytanie, konieczna jest znajomość resztowej wariancji lub przynajmniej orientacyjna wielkość odchylenia standardowego danej cechy w badanych grupach. Przypuśćmy, że w teście  $t$  dla danych zależnych otrzymaliśmy średnią różnicę równą 3.5, odchylenie standardowe dla różnic równe 6.0, a liczebność grupy wynosiła 9. Korzystając z przykładu na s. 27 otrzymamy błąd standardowy różnicy ( $s_{\Delta}$ ) równy  $6.0/\sqrt{9}$ , czyli 2.0. Zatem wartość  $t$  wynosi  $3.5/2 = 1.75$ , co jest wartością nieznamienną ( $P > 0.05$ ). Gdyby przy tych samych wartościach różnicy i SD liczebność wynosiła 14, to  $s_{\Delta}$  wyniosłoby  $6.0/\sqrt{14} = 1.60$ , co daje wartość  $t$  równą  $3.5/1.60 = 2.18$ . Dla  $df = 13$  wartość ta jest znamienna, bowiem  $2.18 > 2.16$  (wartość tablicowa;  $P < 0.05$ ). Analogiczne rozumowanie można zastosować do grup niezależnych bądź do analizy wariancji.

Ponieważ obliczanie wymaganej, czy też pożądanej liczebności grupy zależy od schematu doświadczenia, a także innych czynników, przedstawiono tu jedynie sposób rozumowania, który może być przydatny nawet w złożonych układach.

### **Odrzucanie wątpliwych danych**

Często zdarza się, że jakiś wynik „nie pasuje” do pozostałych – jest w porównaniu z innymi zbyt niski lub zbyt wysoki i ma się ochotę go odrzucić, a często po prostu odrzuca. Tendencja taka jest niebezpieczna, bowiem może prowadzić do odrzucania danych niepasujących do postawionej hipotezy, a nie tylko do rozpatrywanego zbioru.

W takim wypadku należy dokładnie sprawdzić pochodzenie kwestionowanej wartości. Może się np. okazać, że w zbiorze pomiarów siły wykonanych u uczniów jednej klasy znalazła się wartość bardzo niska, pochodząca od ucznia, który właśnie przebył długą chorobę. Taką wartość należy odrzucić dlatego, że pochodzi ona z innej populacji (uczniów, którzy przebyli chorobę), a nie dlatego, że odstaje od innych wartości. Jeżeli zależy nam na charakterystyce zdrowych uczniów, a okaże się, że w tej klasie byli jeszcze inni uczniowie, którzy właśnie przebyli chorobę, to wyniki tych uczniów należy odrzucić, nawet jeżeli mieszczą się w przedziale pozostałych wyników. Reasumując, odrzucenie wyniku powinno



być dokonane na podstawie innych kryteriów niż sama wartość wyniku. Czasem np. okazuje się, że został popełniony tzw. gruby błąd pomiaru, błąd w zapisie wyniku itp. Niekiedy jednak nie udaje się znaleźć takich powodów, a wynik wyraźnie zaburza otrzymany obraz. Można wówczas zastosować kryteria statystyczne, np. 4 odchylenia standardowe od średniej obliczonej bez wartości odstającej, łatwo jednak o popełnienie tzw. błędu I rodzaju (zob. s. 41). O odrzucaniu wątpliwych punktów w badaniu korelacji wspomniano w rozdziale 5.

### Zasady testów statystycznych

Podejmując badania, pragnie się zwykle opisać zjawisko za pomocą obserwacji lub zbadać wpływ jakiegoś czynnika doświadczalnego na interesującą nas cechę. W tym drugim wypadku dąży się zazwyczaj do stwierdzenia, że dany czynnik taki wpływ wywiera. Natomiast w rachunku statystycznym czyni się założenie, że *nie ma* różnicy, *nie ma* wpływu, *nie ma* zależności itp. Założenie takie nazywa się **hipotezą zerową**. Zgodnie z teorią prawdopodobieństwa, hipotezy zerowej nie można ani udowodnić, ani obalić, można jedynie określić jej prawdopodobieństwo. Jeżeli prawdopodobieństwo hipotezy zerowej będzie duże, przyjmiemy tę hipotezę (np. że średnie wartości danej cechy w badanych grupach są podobne), jeżeli małe – odrzucimy (wyciągamy wniosek, że te średnie wartości różnią się znamienne). To, czy prawdopodobieństwo jest duże czy małe, jest wyłącznie sprawą umowną. Za taką krytyczną wartość prawdopodobieństwa przyjęto powszechnie 5%. Ponieważ jest to wartość umowna, nie jest ona bezwzględnie stała i zależnie od sytuacji może być wyższa lub niższa. Na przykład, przy prawdopodobieństwie hipotezy zerowej wynoszącym 10% możemy uznać, że występuje określona tendencja, i przeprowadzić badania uzupełniające.

W zależności od tego, jak zostało sformułowane pytanie badawcze, test statystyczny może być *dwustronny* lub *jednostronny*. Jeżeli w omawianym przykładzie interesuje nas *zmiana* pod wpływem badanego czynnika (treningu), zarówno *in plus*, jak i *in minus*, stosuje się test dwustronny; jeżeli natomiast interesuje nas tylko zmiana w określonym kierunku, np. tylko czy trening jest skuteczny (zob. przykład dla danych zależnych, s. 27), stosuje się test jednostronny. Ewentualna różnica w kierunku odwrotnym niż założony (pogorszenie wyniku testu biegowego) będzie wówczas zignorowana, jako niezgodna z założeniami treningowymi. Oczywiście, jeżeli z hipotezy roboczej wynika, że powinien być zastosowany test dwustronny, to w razie, gdy jego wynik jest nieznamienisty, nie można posłużyć się testem jednostronnym dla zwiększenia czułości.

Aby uściślić pojęcie prawdopodobieństwa hipotezy zerowej, posłużymy się następującym przykładem (ryc. 4.a). Człowiek przychodzący do lekarza może być albo *naprawdę* zdrowy, albo *naprawdę* chory, natomiast lekarz może *stwierdzić* chorobę lub zdrowie. Jeśli u chorego stwierdzi chorobę, a u zdrowego zdrowie, to między stanem faktycznym i orze-

czonym istnieje zgodność. Jeżeli jednak osoba *faktycznie* zdrowa zostanie uznana za chorą, lekarz popełnia **błąd I rodzaju**, w sytuacji zaś, gdy osoba *faktycznie* chora zostanie uznana za zdrową – popełniony jest **błąd II rodzaju**. Ten pierwszy jest oczywiście mniej groźny, bowiem zdrowa osoba uznana za chorą może zostać poddana dalszym badaniom, które powinny doprowadzić do poprawnego rozpoznania. W wypadku błędu II rodzaju, a więc gdy osoba chora zostaje uznana za zdrową, osoba taka traci w istocie szanse na dalsze badania. Otóż, prawdopodobieństwo hipotezy zerowej dotyczy popełnienia błędu I rodzaju, czyli odrzucenia prawdziwej hipotezy.

stan		faktyczny	
		zdrowy	chory
orzeczone	zdrowy	ZGODNE	<b>błąd II rodzaju</b>
	chory	<b>błąd I rodzaju</b>	ZGODNE

**Ryc. 4.a** Schemat ilustrujący pojęcie błędów I i II rodzaju

Błędy II rodzaju, czyli prawdopodobieństwa przyjęcia fałszywej hipotezy, nie da się określić za pomocą rachunków, można jedynie starać się, aby był on jak najmniejszy. Powracając do przykładu medycznego, minimalizacja ryzyka uznania osoby chorej za zdrową wiąże się ze zwiększeniem ryzyka uznania osoby zdrowej za chorą.

Aby można było poprawnie zastosować dany test statystyczny, muszą być spełnione określone warunki. Na przykład, w analizie wariancji (i przy stosowaniu testu  $t$ , który jest jej szczególnym przypadkiem) wymagane jest, aby wariancje w badanych grupach były podobne, a rozkłady wartości – normalne.

Ponieważ normalność rozkładu i jednorodność wariancji są najczęściej spotykanymi wymaganiami, należy je sprawdzić przed przystąpieniem do obróbki danych. Test na jednorodność wariancji (test Bartletta) został omówiony w podrozdziale o jednokierunkowej analizie wariancji (s. 36).

Wyjątkiem od tej zasady są testy oparte na funkcji chi-kwadrat, które są wolne od wstępnych założeń poza koniecznością spełnienia wymogu losowości.

## 5. WSPÓLZALEŻNOŚCI MIĘDZY CECHAMI

Często zadajemy sobie pytanie – czy np. istnieje zależność między barwą oczu a barwą włosów, między siłą dłoni a masą ciała (ryc. 5.a) itp. Wbrew pozorom, są to różne pytania. W pierwszym wypadku, barwa oczu nie może zależeć (w sensie przyczynowym) od barwy włosów, ani też odwrotnie, możemy więc mówić tylko o współzależności między tymi cechami, które przyczynowo mogą zależeć od innej cechy, wspólnej dla obu bądź nie. W drugim wypadku, siła dłoni może zależeć od masy ciała, ale masa ciała od siły dłoni – nie, możemy więc mówić o zależności jednej cechy od drugiej. Nawet jeśli zależność wydaje się logicznie uzasadniona, jak w drugim wypadku, nie musi ona być bezpośrednia, lecz może wynikać np. z silnej zależności między całkowitą masą ciała a masą mięśni kończyny górnej.

Siłę związku (zależności) między dwiema cechami określa **korelacja**, ilościowe zaś zmiany jednej cechy w zależności od zmian drugiej – **regresja**.

### Korelacje dwu zmiennych

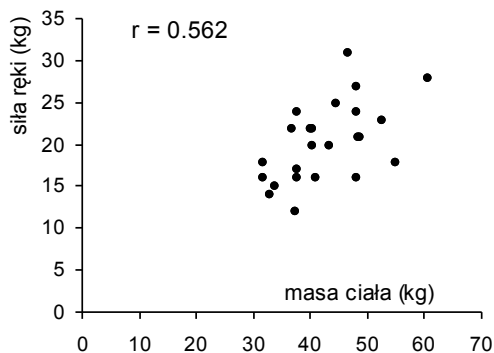
Nie wchodząc w to, czy zależność może mieć charakter przyczynowy czy nie, można określić **siłę współzależności** za pomocą **współczynnika korelacji** ( $r$ ).

$$r = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}} \quad [5-1]$$

We wzorze tym  $S_{xx}$ ,  $S_{yy}$  są zredukowanymi sumami kwadratów zmiennych  $x$ ,  $y$ , a  $S_{xy}$  zredukowaną sumą ich iloczynów. Poniżej przedstawiono zależność między siłą ręki ( $y$ ) a masą ciała ( $x$ ) u 10-letnich chłopców oraz schemat obliczeń.

x	y	x	y
37.1	12	48.0	16
32.9	14	37.6	17
33.7	15	31.7	18
31.7	16	54.9	18
37.5	16	40.3	20
40.7	16	43.3	20
48.2	21	37.6	24
48.5	21	47.9	24
36.5	22	44.3	25
40.0	22	48.1	27
40.2	22	60.4	28
52.3	23	46.6	31

**Tab. 5.1** Masa ciała ( $x$ ) i siła ręki ( $y$ ) u 10-letnich chłopców

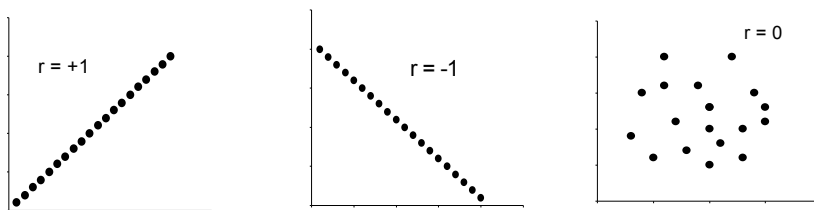


**Ryc. 5.a** Wykres zależności między siłą ręki a masą ciała (dane z tab. 5.1) zwany korelogramem

Jak widać na rycinie 5.a, im większa masa ciała, tym większa na ogół siła ręki, jednak rozrzut jest dość znaczny, a zatem i niezbyt wysoki jest współczynnik korelacji ( $r = 0.562$ ). Rachunki związane z obliczeniem współczynnika korelacji i parametrów regresji pokazano poniżej (dane z tab. 5.1).

$n = 24$ $\sum X = 37.1 + 32.9 + \dots + 46.6 = 1020.0$ $\sum X^2 = 37.1^2 + 32.9^2 + \dots + 46.6^2 = 44660.8$ $\sum xy = 37.1 \cdot 12 + 32.9 \cdot 14 + \dots + 46.6 \cdot 31 = 21202.7$ $\sum y = 12 + 14 + \dots + 31 = 488$ $\sum y^2 = 12^2 + 14^2 + \dots + 31^2 = 10440$ $S_{xx} = \sum X^2 - (\sum X)^2/n = 44660.8 - (1020.0)^2/24 = 1310.8$ $S_{xy} = \sum xy - (\sum X \cdot \sum y)/n = 462.7$ $S_{yy} = \sum y^2 - (\sum y)^2/n = 517.3$	$r^2 = S_{xy}^2 / (S_{xx} \cdot S_{yy}) = 0.316$ $r = 0.562$ $b_{yx} = S_{xy} / S_{xx} = 0.353$ $a = (\sum y - b \cdot \sum X) / n = 5.33$ $B_{yx} = S_{xy}^2 / S_{xx} = 163.3$ $S_{RR} = S_{yy} - B_{yx} = 354.0$ $S_R^2 = S_{RR} / (n-2) = 16.1$ $S_R = 4.01$ $s_0 = S_R / \sqrt{n}$ $Y = 5.33 + 0.353 \cdot X \pm 4.01$
---	--

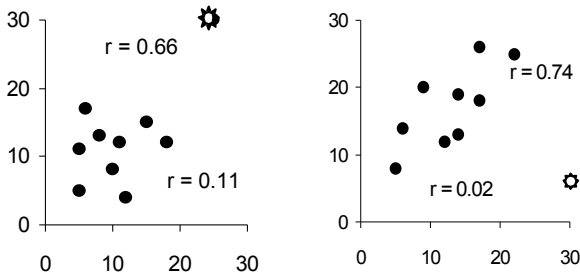
Współczynnik korelacji może przybierać wartości od  $-1$  do  $+1$ , jak pokazano poniżej.



**Ryc. 5.b** Korelogramy, dla których wartości współczynnika korelacji wynoszą  $+1$ ,  $-1$  oraz  $0$

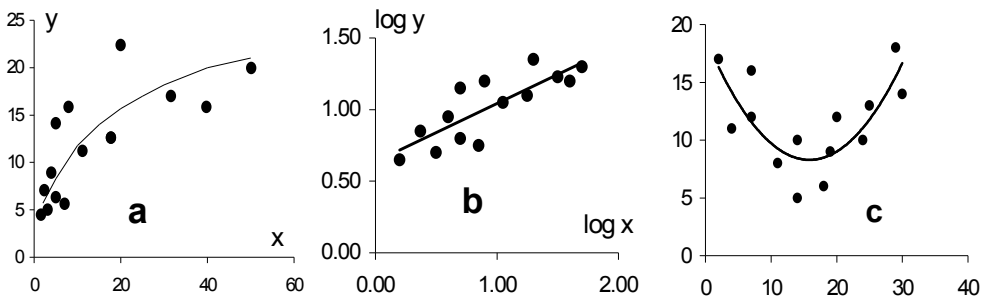
Z samej wielkości współczynnika korelacji nie można jeszcze wnioskować o sile związku, zależy to bowiem także od liczebności próby, a ściślej – od liczby stopni swobody. Na przykład, przy  $n = 8$  ( $df = 6$ ) korelacja równa  $0.70$ , a więc wysoka, nie będzie nawet

znamienna, a przy  $n = 100$  znamienna będzie już tak niska korelacja jak 0.20. Należy ponadto pamiętać o praktycznej regule: jeżeli usunięcie jednego punktu drastycznie zmienia wielkość współczynnika korelacji, to raczej należy polegać na korelacji bez tego punktu (patrz obok; usuwany punkt jest pusty w środku). Jak widać, w pierwszym wypadku usunięcie punktu obniża współczynnik korelacji z 0.66 do 0.11, a w drugim – podwyższa z 0.02 do 0.74. Różnica jest w obu wypadkach zasadnicza.



**Ryc. 5.c** Sytuacje, gdy usunięcie jednego, odstającego punktu (zaznaczonego gwiazdką), drastycznie zmienia wartość współczynnika korelacji

Może się zdarzyć, że korelacja jest wyraźna, ale nieliniowa – jak na wykresach poniżej (ryc. 5.d-a i 5.d-c). W pierwszym wypadku zwykle wystarczy zlogarytmowanie danych, żeby uzyskać zależność liniową (ryc. 5.d-b), natomiast w drugim (5.d-c) – nie oblicza się współczynnika korelacji, tylko tzw. **stosunek korelacyjny** (definicja i obliczanie – patrz s. 57).



**Ryc. 5.d** Korelacja krzywoliniowa (a), która po zlogarytmowaniu danych (b) daje zależność liniową. (c) Korelacja paraboliczna.

Ustalanie siły współzależności za pomocą współczynnika korelacji powinno się w zasadzie stosować wówczas, gdy obie badane zmienne mają rozkład losowy. Gdy jedna zmienna (niezależna,  $x$ ) nie jest losowa, np. gdy ma z góry ustalone wartości – grupy wiekowe w badaniach rozwojowych, obciążenia wysiłkowe itp., współczynnik korelacji zależy od zakresu zmiennej  $x$ , a więc nie mierzy *de facto* współzależności (patrz także rozdział o regresji).

Bardzo ważną wielkością jest kwadrat współczynnika korelacji ( $r^2$ ), zwany **współczynnikiem determinacji**. Określa on, jaka część wariancji zmiennej zależnej (y) jest „wyjaśniana” przez zmienną zależną: w przykładzie na rycinie 5.a  $r = 0.562$ , zatem  $r^2 = 0.316$ , czyli ok. 32% wariancji siły dłoni zależy od masy ciała, a 68% od innych czynników.

*Porównywanie współczynników korelacji:* Ponieważ współczynnik korelacji zawarty jest w przedziale od -1 do +1, jego rozkład jest asymetryczny. Jedynie przy  $r = 0$  rozkład jest normalny. Aby można było porównać dwa współczynniki korelacji pod względem wielkości, należy je najpierw przekształcić na zmienną  $z$  o rozkładzie normalnym w całym zakresie:

$$z = \frac{1}{2} \ln \left( \frac{1+r}{1-r} \right) \quad [5-2]$$

Następnie oblicza się wartość funkcji  $t$  Studenta:

$$t = \Delta z \cdot \sqrt{\frac{m_1 \cdot m_2}{m_1 + m_2}} \quad [5-3]$$

gdzie  $m = n - 3$ , a  $df = \infty$ . Wartości funkcji  $z$  podane są w tablicy 9 (s. 113).

Można również porównywać współczynniki korelacji z tej samej próby, np.  $r_{xy}$  i  $r_{xu}$  lub współczynnik korelacji cząstkowej (patrz dalej) z odpowiednim współczynnikiem korelacji prostej.

Aby zilustrować badanie różnic między współczynnikami korelacji, rozpatrzmy następujące przykłady:

a. Korelacja między siłą dłoni a masą ciała u chłopców ( $n = 24$ ) wyniosła 0.562 (dane z tab. 5.1, patrz s. 42). W podobnej grupie dziewcząt ( $n = 20$ ) stwierdzono  $r = 0.288$ . Korelacja u chłopców była wysoce znamienna ( $P < 0.01$ ), u dziewcząt nie ( $P > 0.10$ ). Zbadano, czy oba współczynniki korelacji różnią się znamienne. W tym celu odczytano w tablicy 9 odpowiednie wartości funkcji  $z$  i podstawiono do wzoru 5-3:  $t = (0.636 - 0.296) \cdot \sqrt{21 \cdot 17 / (21 + 17)} = 1.04$ . Wartość ta jest mniejsza od tablicowej (1.96) dla  $P = 0.05$ , a więc współczynniki korelacji u chłopców i dziewcząt nie różnią się znamienne, mimo że jeden z nich był wysoce znamienny, a drugi nie.

b. W grupie 25 dziewcząt w wieku 9 – 16 lat zmierzono wysokość i masę ciała oraz wydalanie hormonu wzrostu (HGH) z moczem. Otrzymano następującą macierz korelacji:

	h	m	g
a	0.854***	0.629***	-0.250
h		0.755***	0.086
m			-0.216

**Tab. 5.2** Współczynniki korelacji między wiekiem (a), wzrostem (h), masą ciała (m) i wydalaniem hormonu wzrostu (g) u dziewcząt ( $n = 25$ )  
\*\*\*  $P < 0.001$

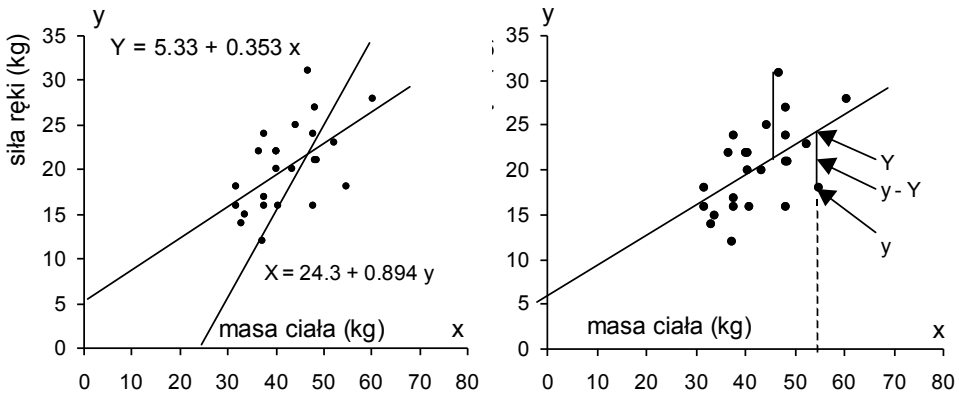
Jak widać, wydalanie HGH nie koreluje znamienne z żadną z pozostałych zmiennych. Jeżeli jednak obliczy się korelacje cząstkowe (wg wzoru 6-5, s. 67) między HGH a wzrostem lub masą ciała z wyłączeniem wpływu wieku, otrzymamy:  $r_{gh \cdot a} = 0.595^{**}$ ,  $r_{gm \cdot a} = -0.078$ . Korelacja cząstkowa z masą ciała pozostaje bliska zeru, natomiast korelacja z wysokością ciała zmienia się na wysoce znamienne. Porównując korelację prostą ( $r_{gh}$ ) z cząstkową przy stałym wieku ( $r_{gh \cdot a}$ ) za pomocą wzoru 5-3 otrzymamy  $t = 1.99$  ( $P < 0.05$ ). Korelacja cząstkowa jest zatem znamienne wyższa od odpowiedniej korelacji prostej. Wynika z tego wniosek, że wydalanie hormonu wzrostu jest związane znamienne nie z masą ciała, a z jego wysokością.

### Regresja liniowa dwu zmiennych

Rachunkowo, regresja związana jest z korelacją, jednak stosuje się ją w innych sytuacjach. Regresja jest ilościową miarą zależności zmiennej  $y$  od zmiennej  $x$ , a więc określa, o ile zmieni się zmienna zależna ( $y$ ), jeżeli zmienna niezależna ( $x$ ) zmieni się o jeden. Zależność tę opisuje **równanie regresji** o postaci

$$\boxed{y = a + bx} \quad , \quad [5-4]$$

gdzie  $a$  – stała regresji, a  $b$  – współczynnik regresji, równy tangensowi kąta nachylenia linii regresji do osi  $x$ . Odpowiedni przykład obliczeń pokazano na s. 43 (przy współczynniku regresji), a wykres regresji – poniżej.



**Ryc. 5.e** Zależność między siłą ręki ( $y$ ) i masą ciała ( $x$ ) u 11-letnich chłopców. (dane z tab. 5.1). Z lewej: regresje  $y$  na  $x$  ( $Y$ ) i  $x$  na  $y$  ( $X$ ); z prawej: regresja  $y$  na  $x$  z zaznaczonymi odległościami dwóch punktów od linii regresji.

$Y$  – punkt na linii regresji odpowiadający zaznaczonej wartości  $x$  (54.9) wyznaczony równaniem ( $Y = 24.7$ );  $y$  – zmierzona wartość siły ręki (18);  $y - Y$  - odległość punktu  $y$  od linii regresji

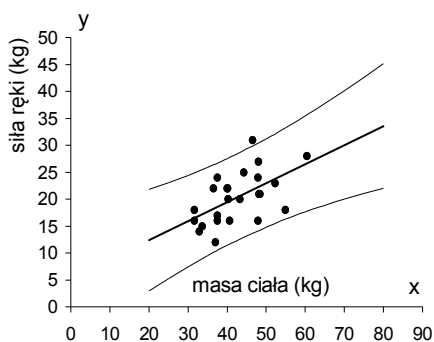
Inaczej niż korelacja, która nie rozróżnia zależności  $y$  od  $x$  oraz  $x$  od  $y$ , regresje dla tych dwu zależności nie są jednakowe. Wynika to z definicji linii regresji, która jest wyznaczana metodą najmniejszych kwadratów: jest to taka linia, dla której suma kwadratów odległości pojedynczych punktów od tej linii, mierzonych *wzdłuż osi  $y$* , jest najmniejsza. Dotyczy to regresji  $y$  na  $x$ , natomiast dla regresji  $x$  na  $y$  odległości punktów od linii regresji będą oczywiście mierzone wzdłuż osi  $x$ . Równanie regresji  $y$  na  $x$  służy zatem do przewidywania wartości zmiennej  $y$  dla danej wartości  $x$ .

Dla omawianego przykładu można zatem obliczyć oczekiwaną siłę ręki znając tylko masę ciała. Wartość 54.9 kg masy ciała (tab. 5.1) podstawiona do równania  $Y = 5.33 + 0.353x$  daje w wyniku 24.9 kg siły ręki. Zmierzona, nie zaś obliczona siła ręki wyniosła jednak dla tego chłopca 18 kg, a więc błąd jest znaczny. Średnie odchylenie pojedynczego punktu od linii regresji ( $s_R$ ) jest odpowiednikiem odchylenia standardowego, ale liczonego nie od punktowej średniej arytmetycznej, a od linii regresji. W tym przykładzie  $s_R = 4.01$  kg (patrz obliczenia na s. 43). Odpowiednikiem błędu standardowego średniej arytmetycznej (SE) jest błąd standardowy w regresji, a więc błąd standardowy  $\bar{x}$  :

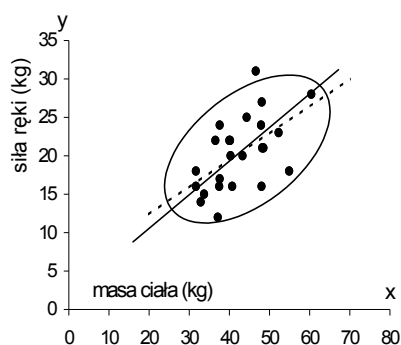
$$s_o = \frac{s_R}{\sqrt{n}} \quad [5-5]$$

Dochodzą tu jeszcze dodatkowe błędy – współczynnika regresji ( $s_b$ ) i stałej regresji ( $s_a$ ). Wzory tych błędów podane są w tabeli 5.3. Z błędu  $s_b$  wynika ważna konsekwencja. Mianowicie, błąd w regresji nie jest stały wzdłuż linii regresji, lecz zwiększa się w miarę oddalania się od środka regresji. Ów błąd dany jest wzorem 5-6, a wykres przedstawiający linię regresji i linie podwójnego błędu w regresji – na rycinie 5.f.

$$s_y^2 = s_R^2 + s_b^2 \cdot (x - \bar{x})^2 \quad [5-6]$$



**Ryc. 5.f** Równanie regresji dla omawianego przykładu z liniami wyznaczającymi podwójny błąd w regresji:  $Y = 5.33 + 0.353x \pm 2s_y$ . Linie błędów ekstrapolowano poza zakres zmiennej  $x$ , aby pokazać ich zakrzywienie.



**Ryc. 5.g** Elipsa korelacyjna, w której mieszczą się punkty odpowiadające wartościom zmiennych  $x$  i  $y$ . Długa oś elipsy nie pokrywa się z linią regresji.



Obszar określony regresją, a ściślej krzywymi błędu (ryc. 5.f) nie pokrywa się z eliptycznym obszarem korelacyjnym (ryc. 5.g), co wynika z innego ujęcia danych przez regresję w porównaniu z korelacją.

Należy pamiętać, że równanie regresji można stosować tylko dla tego zakresu wartości  $x$ , dla którego wyznaczono regresję, a więc nie należy stosować ekstrapolacji. Można to uzasadnić podanym przykładem: w wyniku ekstrapolacji regresji do zerowej masy ciała, nie otrzyma się zgodnie z logiką zerowej wartości siły ręki, lecz wartość 5.33 kg (stała regresji). Obliczone równanie liniowe jest więc tylko przybliżeniem użytecznym w danym przedziale wartości  $x$  i nie należy go traktować jako opisu zjawiska. Ewentualny opis musiałby spełniać warunek przejścia linii regresji przez początek układu współrzędnych, a więc przez punkt 0,0. Można to osiągnąć narzucając równanie  $Y = bx$  ( $a = 0$ ), jak podano w tabeli 5.3 lub obliczając liniową regresję po uprzednim zlogarytmowaniu wartości  $y$ , a ew. także wartości  $x$  (zob. s. 52).

Jak wspomniano wyżej, niekiedy zachodzi konieczność obliczenia linii regresji przy określonych założeniach, np. że musi ona przejść przez zero, że współczynnik regresji musi być równy jedności itp. Wzory na obliczenie parametrów i błędów niektórych regresji podaje poniższa tabelka.

Równanie	$b =$	$a =$	$S_{RR} =$	$s_b^2 =$	$s_a^2 =$
$Y = a + bx$	$S_{xy}/S_{xx}$	$(\sum y - b \cdot \sum x)/n$	$S_{yy} - S_{xy}^2/S_{xx}$	$s_R^2/S_{xx}$	$s_b^2 \cdot \sum x^2/n$
$Y = a + x$	1	$(\sum y - \sum x)/n$	$S_{yy} - 2S_{xy} + S_{xx}$	$s_R^2/S_{xx}$	$(s_R^2/n) + s_b^2 \cdot \bar{x}^2$
$Y = bx$	$\sum xy/\sum x^2$	0	$\sum y^2 - (\sum xy)^2/\sum x^2$	$s_R^2/S_{xx}$	$(s_R^2/n) + s_b^2 \cdot \bar{x}^2$
$Y = x$	1	0	$\sum y^2 + \sum x^2 - 2\sum xy$	$s_R^2/S_{xx}$	$(s_R^2/n) + s_b^2 \cdot \bar{x}^2$

**Tab. 5.3** Wzory na parametry regresji liniowej spełniającej zadane warunki. Dla każdego typu regresji błąd w regresji jest równy:  $s_R = \sqrt{S_{RR}/(n-2)}$

Przedstawiony tu przykład odnosi się do tzw. regresji II rodzaju, tj. gdy obie zmienne są losowe. W takim wypadku wybór zmiennej niezależnej może być dowolny. Gdy jedna ze zmiennych jest nielosowa, wówczas nie może ona być zmienną zależną; jest to tzw. regresja I rodzaju. Przykładem może być krzywa kalibracyjna w analizie laboratoryjnej, tam bowiem wartości zmiennej niezależnej (np. stężenia wzorcowe) są celowo dobrane. Jak wspomniano w omówieniu korelacji dwu zmiennych, dla zależności I rodzaju nie oblicza się współczynnika korelacji. Właściwą miarą jest tu błąd w regresji.

### Analiza regresji liniowej

W tabeli 5.4 przedstawiono pomiary wysokości ciała ( $w$ ) i długości tułowia ( $t$ ; „wysokość na siedząco”) w dwu grupach dziewcząt – z zespołem Turnera i zdrowych (grupa

kontrolna), dobranych do poprzednich pod względem wieku i wzrostu. Interesuje nas zbadanie proporcji wzrostowych w obu grupach dziewcząt. Najprostszym wyjściem wydaje się obliczenie stosunku długości tułowia do wysokości ciała ( $t/w$ ) i porównanie średnich wartości, ale spróbujmy zanalizować zagadnienie nieco dokładniej za pomocą zależności między  $t$  i  $w$  (ryc. 5.h).

Grupa kontrolna			Zespół Turnera		
w	t	t/w	w	t	t/w
129.5	69.9	0.540	129.6	69.0	0.532
137.1	72.1	0.526	130.9	72.2	0.552
137.5	69.9	0.508	131.1	71.1	0.542
138.9	72.2	0.520	133.4	74.5	0.558
139.5	73.4	0.526	134.2	72.6	0.541
141.5	74.6	0.527	135.6	77.2	0.569
141.9	74.7	0.526	135.9	79.2	0.583
141.9	76.2	0.537	136.7	75.4	0.552
142.2	77.5	0.545	139.0	74.6	0.537
142.4	72.9	0.512	140.4	76.2	0.543
142.9	74.9	0.524	141.3	76.0	0.538
143.1	74.8	0.523	150.9	81.0	0.537
143.1	74.8	0.523	148.0	79.5	0.537
143.1	72.6	0.507			
143.9	74.3	0.516			
148.3	77.4	0.522			
141.1	73.9	0.524	137.5	75.3	0.548
4.0	2.2	0.010	6.2	3.3	0.014

**Tab. 5.4** Wysokość ciała ( $w$ ), długość tułowia ( $t$ ) oraz stosunek  $t/w$  u dziewcząt zdrowych ( $n = 16$ ) i z zespołem Turnera ( $n = 13$ ). W ostatnich dwóch wierszach podano średnie i odchylenia standardowe.

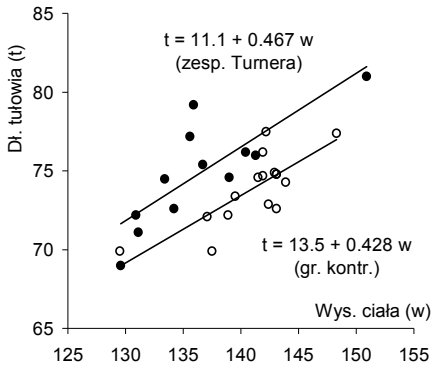
Obliczenia wykonuje się w podobny sposób jak dla regresji liniowej (s. 53), dlatego w tabelach na następnej stronie podano tylko wielkości poszczególnych składowych analizy.

Z wartości zestawionych w tabeli 5.5 oblicza się odpowiednie zredukowane sumy jak w przykładzie na s. 43 i zestawia w tabelę analizy regresji (tab. 5.6).

W analizie regresji otrzymuje się dwie wartości testu  $F$ : dla zróżnicowania między współczynnikami regresji [ $b$ ] oraz między stałymi regresji [ $a$ ]. W omawianym przykładzie współczynniki regresji dla obu grup nie różnią się znamienne ( $F < 1$ ), dlatego obliczono ‘wspólny’ współczynnik regresji:

$$b_{\text{wsp.}} = 331.08/746.44 = 0.4435$$

oraz resztową wariancję w regresji:  $s_R^2 = 74.22/26 = 2.855$ .



**Ryc. 5.h** Wykresy danych z tabeli 5.4 z naniesionymi liniami i równaniami regresji

	Grupa kontrolna	Zespół Turnera	Razem
n	16	13	29
$\Sigma w$	2256.8	1787.0	4043.8
$\Sigma t$	1182.2	978.5	2160.7
$\Sigma w^2$	318571.8	246140.1	564711.9
$\Sigma wt$	166856.4	134730.2	301586.6
$\Sigma t^2$	87425.3	73796.6	161221.8

**Tab. 5.5** Podstawowe wielkości (sumy, sumy kwadratów, sumy iloczynów) potrzebne do analizy regresji obliczone jak w przykładzie na s. 42

	$S_{ww}$	$S_{wt}$	$S_{tt}$	df	$S_{RR}$	df	$s_R^2$	F
GK	250.16	107.05	75.46	15	29.65	14	$b = 0.4279$	
ZT	496.28	224.03	145.60	12	44.47	11	$b = 0.4514$	
$\Sigma$	746.44	331.08	221.06	27	<b>74.22</b>	26	2.855	
[b]	wspólne $b = 0.4435$				74.12	25		
Razem	838.75	295.50	234.72	28	130.61	27		
[a]					56.39	1	56.39	19.8***

**Tab. 5.6** Tablica analizy regresji dla danych z tabel 5.4 i 5.5; GK – grupa kontrolna, ZT – zespół Turnera,  $\Sigma$  – sumy dwóch poprzednich wierszy, Razem – sumy resztowe obliczone z kolumny ‘Razem’ w tabeli 5.5. Wytluszczona liczba w wierszu ‘ $\Sigma$ ’ jest wartością  $S_{RR}$  obliczoną z sum wiersza ‘ $\Sigma$ ’, natomiast wartość poniżej (74.12) jest sumą wartości  $S_{RR}$  z wierszy GK i ZT.

Zróznicowanie między stałymi regresji jest natomiast bardzo wysoce znamienne ( $F = 19.8$ ;  $P < 0.001$ ). Wyniki te oznaczają, że badane linie regresji są równoległe (brak zróżnicowania w  $b$ ) i znamienne przesunięte względem siebie. Stałe regresji dla obu grup oblicza się jak na s. 43, podstawiając do wzoru wspólne  $b$ :

$$a_{GK} = (1182.2 - 0.4435 \cdot 2256.8/16) = 11.33 \pm 15.07$$

$$a_{ZT} = (978.5 - 0.4435 \cdot 1787.0/13) = 15.85 \pm 10.44$$

Podstawiając odpowiednie dane do wzoru z tabeli 5.3 oblicza się podane wyżej błędy stałych regresji. Jak widać, żadna ze stałych nie jest statystycznie różna od zera, co wynika z wysokiej średniej wysokości ciała w stosunku do obserwowanego zakresu.

Gdyby zróżnicowanie między współczynnikami regresji ( $b$ ) było statystycznie istotne, należałoby przyjąć równania regresji obliczone dla każdej grupy oddzielnie.

Warunkiem poprawności analizy regresji jest brak zróżnicowania między badanymi grupami pod względem średniej wartości zmiennej  $x$ . Jeżeli bowiem rozkłady zmiennej  $x$  są w badanych grupach normalne, to różnice między średnimi wynikają z różnych zakresów tej zmiennej. Ponieważ zaś regresja obowiązuje tylko w obrębie zakresu zmiennej  $x$ , porównanie regresji o różnych zakresach wymagałoby ekstrapolacji (zob. s. 47).

W tabeli 5.4 obok wartości  $w$  i  $t$  podano również stosunek  $t/w$ . Zamiast przeprowadzać analizę regresji można by obliczyć statystyczną różnicę między średnimi wartościami  $t/w$ , jak podano na s. 26. Podstawiając odpowiednie wartości do wzoru 4-1 otrzymamy  $\Delta = 0.024$ ,  $s_{\Delta} = 0.0045$ , zatem  $t = 5.33$  ( $P < 0.001$ ). Należy jednak pamiętać, że posługiwanie się stosunkiem zmiennej  $y$  do  $x$  ma sens tylko wówczas, gdy korelacja między tymi zmiennymi jest dodatnia i liniowa, a linia regresji przechodzi przez punkt 0,0 (stała regresji nie różni się statystycznie od zera). Celem posługiwania się liczbami względnymi jest redukcja wariancji zmiennej  $y$  wywołanej wpływem zmiennej niezależnej ( $x$ ). Przy zerowej korelacji zmienność  $y/x$  nie będzie mniejsza niż zmienność  $y$ , a przy korelacji ujemnej – będzie nawet większa. Z drugiej strony, jeżeli regresja  $y$  na  $x$  nie przechodzi przez początek układu współrzędnych, wówczas wartości  $y/x$  będą w dalszym ciągu skorelowane z wartościami  $x$ . Zgodnie z logiką, linie regresji dla obu grup powinny wychodzić z początku układu współrzędnych (punkt 0,0), ale wówczas będą w przybliżeniu równoległe tylko w małym zakresie zmiennej  $x$ . Wszystkich tych niedogodności można uniknąć przez zlogarytmowanie wartości obu zmiennych (patrz s. 54). Tu ograniczamy się jedynie do zwrócenia uwagi na tę możliwość i do omówienia różnych aspektów analizy regresji.

### Regresje krzywoliniowe

Zależności między dwiema cechami są bardzo często nieliniowe, przy czym nieliniowość może mieć różną postać. Można wyróżnić 3 główne typy funkcji nieliniowych: *monotoniczne* (rosnące lub malejące), *minimaksowe* (zawierające jedno lub więcej minimów i/lub maksimów) i *cykliczne* (okresowe). Przykłady takich krzywych pokazano na rycinie 5.i.



**Ryc. 5.i** Różne zależności nieliniowe: (a) zależność monotonicznie rosnąca przekształcona w postać liniową (b) przez zlogarytmowanie danych; (c) zależność minimaksowa, paraboliczna; (d) zależność cykliczna, okresowa

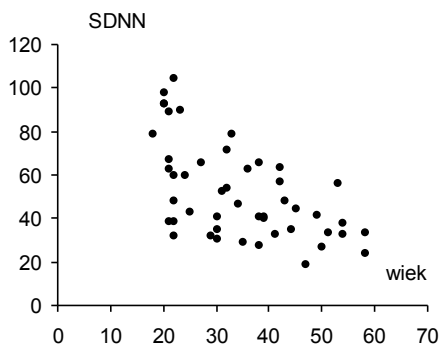
Analiza takich zależności polega na odpowiednim przekształceniu (transformacji) jednej lub obu zmiennych tak, aby otrzymać zależność liniową (z zachowaniem warunku stałości wariancji, tzn. w przybliżeniu jednakowego rozrzutu punktów wzdłuż całej linii zależności). Sposób ten dotyczy jednak tylko zależności monotonicznych. Do krzywych niemonotonicznych należy dopasować odpowiednią funkcję matematyczną, możliwie dobrze opisującą badany zbiór. Do krzywych minimaksowych stosuje się najczęściej tzw. wielomiany ortogonalne, czyli parabole odpowiedniego stopnia, przy czym ograniczymy się tu do omówienia regresji drugiego stopnia (kwadratowej). Krzywe cykliczne przybliża się funkcją trygonometryczną (kosinus), a właściwie kombinacją kosinusów o różnych okresach (o różnej częstotliwości). Analiza krzywych cyklicznych, jak również innych postaci krzywych nie będzie tu omawiana, jest to bowiem zbyt specjalistyczne zagadnienie.

#### *Przekształcenie regresji w postać liniową*

Rozpatrzmy następujący przykład: w grupie 48 zdrowych, nietreningujących osób w wieku od 18 do 53 lat wykonano pomiary tzw. całkowitej zmienności rytmu serca (SDNN) przedstawione w tabeli 5.7.

wiek	SDNN	wiek	SDNN	wiek	SDNN	wiek	SDNN	wiek	SDNN
18	79	22	48	30	31	38	28	49	42
20	98	22	39	31	53	39	41	50	27
20	93	22	32	32	72	39	40	51	34
20	93	23	90	32	54	41	33	54	38
21	89	24	60	33	79	42	64	54	33
21	67	25	43	34	47	42	57	58	34
21	63	27	66	35	29	43	48	58	24
21	39	29	32	36	63	44	35	53	56
22	105	30	41	38	66	45	45		
22	60	30	35	38	41	47	19		

**Tab. 5.7** Wartości wskaźnika zmienności rytmu serca (SDNN) u zdrowych osób w różnym wieku

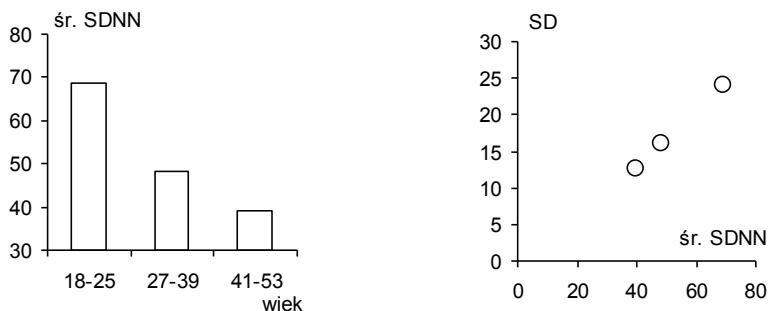


**Ryc. 5.j** Zależność między wskaźnikiem zmienności rytmu serca (SDNN) a wiekiem. Widoczne obniżanie się średniej i rozrzutu z wiekiem.

wiek	$\bar{x}$	SD
18 – 25	68.6	24.1
27 – 39	48.1	16.1
41 – 53	39.3	12.7

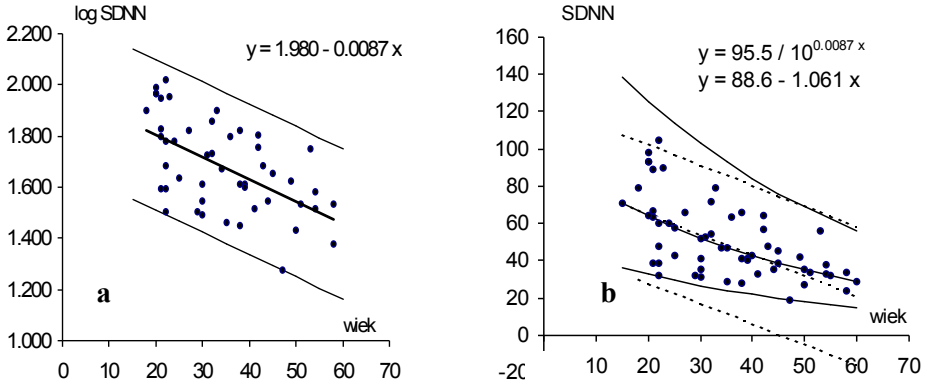
**Tab. 5.8** Średnie wartości i odchylenia standardowe SDNN w 3 grupach wiekowych (z danych w tab. 5.7)

Na wykresie zależności SDNN od wieku przedstawionym na rycinie 5.j widać, że wartości te obniżają się z wiekiem przy jednoczesnym zmniejszaniu się rozrzutu. Aby zyskać przybliżony obraz zbiorczy tej zależności, podzielmy dane na 3 grupy wiekowe i obliczmy dla każdej grupy średnią i odchylenie standardowe wartości SDNN (zob. tab. 5.8).



**Ryc. 5.k** Średnie wartości SDNN w 3 grupach wiekowych (z lewej) i zależność odchylenia standardowego SDNN od średniej (z prawej). Według danych z tabeli 5.8.

Zmiany średniej wartości SDNN z klasą wieku oraz zmiany odchylenia standardowego (SD) w zależności od średniej pokazano na rycinie 5.k. Zarówno krzywoliniowy spadek SDNN z wiekiem, jak i proporcjonalność SD do średniej SDNN są przesłanką do logarytmowania wartości SDNN.



**Ryc. 5.1** (a) Regresja liniowa dla danych logarytmicznych i linie odległe od regresji o  $\pm 2$  SD; (b) Regresja liniowa dla danych surowych (linie kropkowane) i regresja otrzymana ze zdelogarytmowania regresji (a)

Na rycinie 5.1 (z lewej) pokazano regresję wartości logarytmicznych (nie podanych w tabeli) i linie oddalone o dwa odchylenia standardowe od regresji. Jak widać, rozrzut punktów jest w przybliżeniu jednakowy wzdłuż linii regresji. Z prawej strony pokazano linię regresji (kropkowaną) obliczoną z danych surowych (nie logarytmowanych) oraz kropkowane linie oddalone o 2 SD od regresji, liniami ciągłymi zaś – linię otrzymaną ze zdelogarytmowania regresji logarytmicznej oraz analogiczne linie oddalone o 2 SD od regresji. Logarytmowanie danych pozwoliło zatem uzyskać statystyczny opis zależności zgodny z obserwowanym obrazem.

#### *Regresja paraboliczna drugiego stopnia (kwadratowa)*

Jeżeli dany zbiór ma jedno maksimum (lub minimum), sprowadza się do znalezienia równania kwadratowego o postaci

$$\boxed{y = b_0 + b_1 \cdot x + b_2 \cdot x^2} \quad [5-7]$$

Przykład takiego zagadnienia przedstawiono poniżej.

RE	PS	log PS	RE	PS	log PS
0	37.7	1.576	10	29.3	1.467
2	20.5	1.311	11	13.5	1.129
2	29.5	1.470	12	16.5	1.217
4	13.3	1.123	12	39.8	1.600
4	19.0	1.278	13	22.4	1.350
5	21.7	1.337	13	35.2	1.547
6	23.0	1.361	14	17.5	1.244
7	10.3	1.014	14	24.4	1.387
7	15.7	1.196	14	68.8	1.838
9	11.3	1.052	15	21.6	1.334
10	25.2	1.401	17	32.5	1.512

**Tab. 5.9** Reaktywność emocjonalna (RE) i wielkość skórno przepływu krwi (PS) oraz jego logarytmu (log PS) u zdrowych osób w spoczynku

W tabeli 5.9 przedstawiono pomiary skórno przepływu krwi – bezpośrednie (PS) i zlogarytmowane (log PS) oraz reaktywności emocjonalnej, a na rycinie 5.m – wykresy zależności. Zależność ta jest dla surowych danych wyraźnie paraboliczna, ale im wyższa reaktywność, tym większe rozrzuty punktów. Dlatego wartości przepływu zlogarytmowano, co w znacznym stopniu wyrównało ów rozrzut. Obliczenia przeprowadza się następująco:

$n = 22$ $\Sigma x = 0+2+ \dots +17 = 201$ $\Sigma x^2 = 0^2+2^2+ \dots +17^2 = 2329$ $\Sigma x^3 = 0^3+2^3+ \dots +17^3 = 29601$ $\Sigma x^4 = 0^4+2^4+ \dots +17^4 = 396457$ $\Sigma y = 37.7+20.5+ \dots +32.5 = 548.7$ $\Sigma y^2 = 37.7^2+20.5^2+ \dots +32.5^2 = 17160.6$ $\Sigma xy = 0 \cdot 37.7+2 \cdot 20.5+ \dots +17 \cdot 32.5 = 5303.6$ $\Sigma x^2y = 0^2 \cdot 37.7+2^2 \cdot 20.5+ \dots +17^2 \cdot 32.5 = 65151.4$	$S_{xx} = 2329 - 201^2/22 = 492.59$ $S_{x^3} = 29601 - 201 \cdot 2329/22 = 8322.41$ $S_{x^4} = 396457 - 2329^2/22 = 149901$ $S_{xy} = 5303.6 - 201 \cdot 548.7/22 = 290.48$ $S_{x^2y} = 65151.4 - 2329 \cdot 548.7/22 = 7064.02$ $S_{yy} = 17160.6 - 548.7^2/22 = 3475.49$
---	---

Z powyższych wielkości wyznacza się następnie współczynniki  $b_2$ ,  $b_1$  i  $b_0$ , (wzory 5-8) otrzymując równanie (ryc. 5.m):  $Y = 30.81 - 3.331x + 0.2321x^2$ .

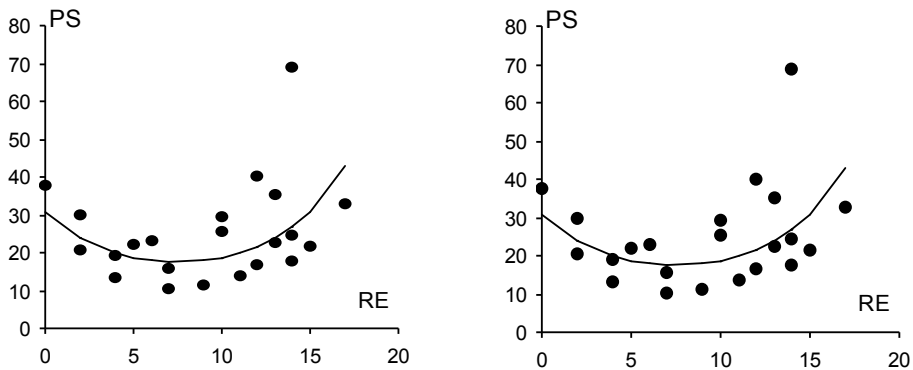
$$b_2 = \frac{S_{xx} \cdot S_{x^2y} - S_{x^3} \cdot S_{xy}}{S_{xx} \cdot S_{x^4} - S_{x^3}^2}$$

$$b_1 = \frac{S_{xy} \cdot S_{x^4} - S_{x^3} \cdot S_{x^2y}}{S_{xx} \cdot S_{x^4} - S_{x^3}^2}$$

$$b_0 = \frac{\Sigma y - b_1 \cdot \Sigma x - b_2 \cdot \Sigma x^2}{n}$$

[5-8]





**Ryc. 5.m** Paraboliczna zależność skórniego przepływu krwi (PS) lub jego logarytmu (log PS) od poziomu reaktywności emocjonalnej (RE)

Analogicznie oblicza się parametry równania dla danych zlogarytmowanych, otrzymując równanie przedstawione na rycinie 5.m:

$$\log Y = 1.486 - 0.06446x + 0.00430x^2.$$

Należy jeszcze obliczyć niektóre inne parametry – błąd w regresji i stosunek korelacyjny. Błąd w regresji oblicza się analogicznie jak dla regresji liniowej trzech zmiennych (s. 77 i 78):

$S_{RR} = S_{yy} - B$ , gdzie  $B = b_1 \cdot S_{xy} + b_2 \cdot S_{x^2y}$ . Po podstawieniu do tego wzoru obliczonych uprzednio wartości otrzymamy:

$$B = 3.331 \cdot 290.48 + 0.2321 + 7064.02 = 2607.15$$

$$S_{RR} = 3475.49 - 2607.15 = 868.34, \quad s_R^2 = 868.34/19 = 45.70, \quad s_R = \sqrt{45.70} = 6.76.$$

Wreszcie, trzeba sprawdzić, czy regresja kwadratowa znamienne różni się od liniowej. W podanym przykładzie paraboliczny przebieg jest oczywisty, ale nie zawsze tak jest. W celu przeprowadzenia tego testu należy obliczyć wartości błędów w regresji liniowej (z danych na poprzedniej stronie) i parabolicznej (powyżej). Wyniki przedstawia poniższa tabela.

Regresja	Obliczanie $S_{RR}$	$S_{RR}$	df	$s^2$	F
Liniowa	$3475.49 - 290.48^2/492.59$	3304.19	20		
Paraboliczna	podane w tekście	868.34	19	45.70	
Różnica	–	2435.85	1	2435.85	53.3***

Miarą siły związku między badanymi cechami nie będzie oczywiście współczynnik korelacji liniowej, bo zależność jest paraboliczna. W takim wypadku, jak również wszędzie tam, gdzie zależność nie jest liniowa, oblicza się tzw. **stosunek korelacyjny**. Jest to

względny udział zmienności związanej z regresją do całkowitej zmienności zmiennej  $y$ , czyli stosunek  $B$  do  $S_{yy}$ :

$$R^2 = \frac{B}{S_{yy}} \quad [5-9]$$

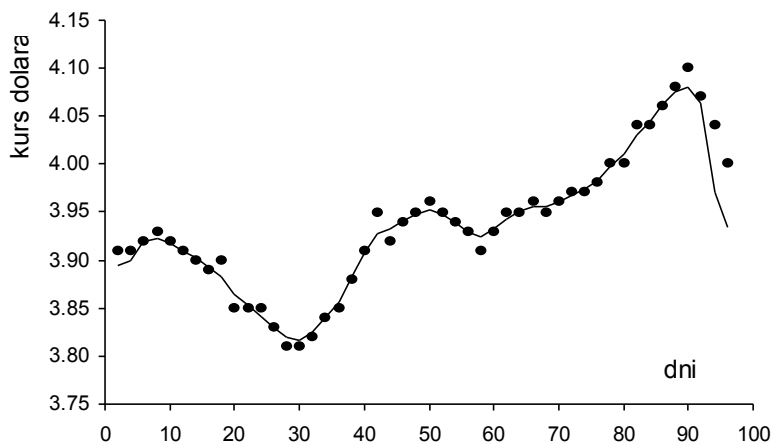
Pierwiastek ze stosunku korelacyjnego ( $R$ ) jest pojęciowo i liczbowo równoważny współczynnikowi korelacji liniowej, ale nie ma określonego znaku. Stosunek korelacyjny jest natomiast identyczny z kwadratem współczynnika korelacji wielokrotnej omówionym w dalszej części (s. 78).

### *Wyrównywanie średnią ruchomą*

Niekiedy zależność nie ma postaci chmury punktów, lecz jest ciągiem pojedynczych wartości np. w czasie. Jako przykład mogą służyć notowania kursu waluty w ciągu pewnego okresu przedstawione w tabeli 5.10. Notowania te, pokazane na rycinie 5.n jako kropki, wykazują nieregularny przebieg, do którego trudno dopasować funkcję matematyczną. Można jednak nieco wygładzić przebieg punktów za pomocą tzw. średniej ruchomej.

3.91	3.91	3.92	3.93	3.92	3.91	3.90	3.89	3.90	3.85	3.85	3.85	3.83	3.81	3.81	3.82
3.84	3.85	3.88	3.91	3.95	3.92	3.94	3.95	3.96	3.95	3.94	3.93	3.91	3.93	3.95	3.95
3.3.9	3.95	3.96	3.97	3.97	3.98	4.00	4.00	4.04	4.04	4.06	4.08	4.10	4.07	4.04	4.00

**Tab. 5.10** Ciągłe notowania kursu dolara (co drugi dzień); po wartościach 3.91 i 4.00 podano tylko liczbę groszy



**Ryc. 5.n** Notowania kursu dolara (kropki; dane z tab. 5.10) i linia wygładzona średnią ruchomą

Poszczególne wartości z serii wyrównuje się średnią ruchomą w ten sposób, że do obliczeń bierze się 5 kolejnych punktów. Jeżeli oznaczy się kolejne wartości jako  $x_1, x_2, x_3, x_4, x_5, x_6$  itd., to poprawiona wartość

$$X_3 = (x_1 + 2x_2 + 4x_3 + 2x_4 + x_5)/10, \quad X_4 = (x_2 + 2x_3 + 4x_4 + 2x_5 + x_6)/10 \text{ itd.}$$

Od razu widać, że nie da się obliczyć wartości  $X_1, X_2$  (a także dwu ostatnich), bowiem nie ma danych, które należy wstawić do wzoru. Wartości te oblicza się następująco:

$$X_2 = (2x_1 + x_2 - x_4)/2, \quad X_1 = (2X_2 + x_1 - x_3)/2$$

Na przykład, średnia ruchoma dla wartości  $x_3$  z tabeli 5.10 wynosi:

$$X_3 = (3.91 + 2 \cdot 3.91 + 4 \cdot 3.92 + 2 \cdot 3.93 + 3.92)/10 = 3.919, \text{ dla wartości } x_2 \text{ i } x_1:$$

$$X_2 = (2 \cdot 3.91 + 3.91 - 3.93)/2 = 3.900, \quad X_1 = (2 \cdot 3.900 + 3.91 - 3.92)/2 = 3.895$$

Analogicznie oblicza się średnie ruchome dla dwu ostatnich wartości. Średnie ruchome obliczone dla całej serii danych z tabeli 5.10 pokazano na rycinie 5.n w postaci linii ciągłej. Jak widać, wyrównanie średnią ruchomą daje wygładzony, uśredniony przebieg w porównaniu z surowymi danymi.

## 6. ANALIZA WIELOCECHOWA

Analiza wielu cech jednocześnie może nastroczać znaczne trudności, bowiem poszczególne zmienne mogą sugerować różne, nawet sprzeczne wnioski. Trudności takie napotyka np. lekarze w różnicowaniu podobnych zespołów chorobowych, trenerzy, mający za zadanie dokonać selekcji zawodników na podstawie wyników kilku testów, lub badacze pragnący porównać kilka grup ze względu na wiele cech. Poniżej zostaną omówione niektóre problemy związane z analizą wielo cechową.

### Profile wielo cechowe

Graficzne przedstawienie średnich wartości kilku cech dla dwu lub więcej grup pozwala na znacznie łatwiejszą ocenę ich wzajemnej konfiguracji niż przegląd tych samych danych ujętych w tabelę. Różne skale zmiennych mogą jednak uniemożliwić pokazanie wszystkich danych na jednym wykresie. Najlepszym sposobem sprowadzenia różnych zmiennych do jednej skali jest ich unormowanie, przedstawione i omówione w rozdziale 3.

Górna część tabeli 6.1 zawiera dane od 7 zawodników uzyskane w teście Wingate, dolna zaś – te same dane po unormowaniu na ich średnie i odchylenia standardowe. Na wykresie (ryc. 6.a) zaznaczono przerywaną linią profile najbardziej podobne do siebie (nr 1, 2, 5, 6), tak więc nawet wizualnie można klasyfikować profile.

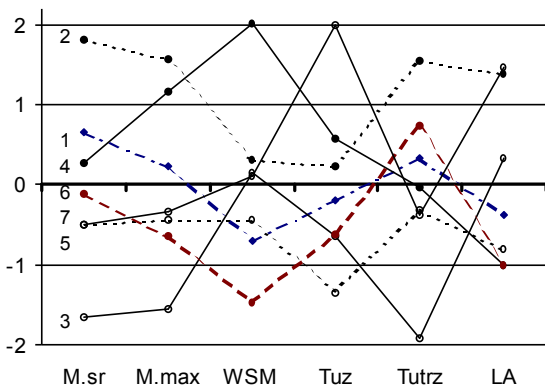
nr	$M_{sr}$	$M_{max}$	WSM	$T_{uz}$	$T_{utr}$	LA	
1	10.3	11.74	0.266	4.24	3.76	9.3	
2	10.9	12.48	0.305	4.43	4.24	11.8	
3	9.1	10.76	0.299	4.04	2.87	10.3	
4	10.1	12.25	0.371	4.58	3.61	8.4	
5	9.7	11.37	0.276	3.72	3.5	8.7	
6	9.9	11.26	0.236	4.05	3.92	8.4	
7	9.7	11.43	0.297	5.21	3.48	11.9	
$\bar{x} \pm SD$	9.96±0.52	11.61±0.55	0.29±0.04	4.32±0.45	3.63±0.39	9.83±1.42	
nr	$M_{sr}$	$M_{max}$	WSM	$T_{uz}$	$T_{utr}$	LA	średnia
1	0.66	0.23	-0.69	-0.19	0.34	-0.37	0.00
2	1.81	1.58	0.31	0.24	1.56	1.39	1.15
3	-1.65	-1.55	0.16	-0.64	-1.91	0.33	-0.88
4	0.27	1.16	2.01	0.57	-0.04	-1.01	0.50
5	-0.49	-0.44	-0.43	-1.36	-0.32	-0.80	-0.64
6	-0.11	-0.64	-1.46	-0.62	0.75	-1.01	-0.52
7	-0.49	-0.33	0.11	1.99	-0.37	1.46	0.39

**Tab. 6.1** Wartości zarejestrowane u 7 zawodników w teście Wingate: moc średnia ( $M_{sr}$ ), moc maksymalna ( $M_{max}$ ), wskaźnik spadku mocy (WSM), czas uzyskania ( $T_{uz}$ ) i utrzymania ( $T_{utr}$ ) mocy maksymalnej, stężenie mleczanu we krwi po teście (LA).

Zmienne unormowane mogą też posłużyć do sporządzenia rankingu pod względem nie jednej, a wielu badanych cech. Najprostszym sposobem jest obliczenie dla każdej osoby średniej wartości zmiennych unormowanych (ostatnia kolumna w tabeli 6.1, dolna część). Należy jednak pamiętać, że zawodnik będzie tym lepszy, im **wyższe** wartości osiągnie w mocy średniej ( $M_{sr}$ ), maksymalnej ( $M_{max}$ ) i w czasie utrzymania mocy maksymalnej ( $T_{utr}$ ), a im **niższe** – we wskaźniku spadku mocy maksymalnej (WSM), czasie uzyskania mocy maksymalnej ( $T_{uz}$ ) i w stężeniu mleczanu (LA). Dlatego, należy zmienić znaki unormowanych wartości tych trzech zmiennych (zaciemnione pola w tabeli 6.2). Średnie wartości zmiennych unormowanych mogą teraz posłużyć do rankingu zawodników. Jak widać, trzech uzyskało średnie wartości ujemne (nr 3, 4, 7), a pozostali – dodatnie. Ci ostatni (nr 1, 2, 5, 6) mieli podobne profile „surowych” wartości unormowanych.

nr	$M_{sr}$	$M_{max}$	-WSM	$-T_{uz}$	$T_{utr}$	-LA	śr.
1	0.66	0.23	0.69	0.19	0.34	0.37	0.41
2	1.81	1.58	-0.31	-0.24	1.56	-1.39	0.50
3	-1.65	-1.55	-0.16	0.64	-1.91	-0.33	-0.83
4	0.27	1.16	-2.01	-0.57	-0.04	1.01	-0.03
5	-0.49	-0.44	0.43	1.36	-0.32	0.80	0.22
6	-0.11	-0.64	1.46	0.62	0.75	1.01	0.51
7	-0.49	-0.33	-0.11	-1.99	-0.37	-1.46	-0.79

**Tab. 6.2** Te same dane co w dolnej części tabeli 6.1, ale dla unormowanych wartości WSM,  $T_{uz}$  i LA zmieniono znaki, bowiem dla tych cech korzystne są wartości jak najniższe



**Ryc. 6.a** Profile wielocechowe 7 zawodników wykreślone z danych w tabeli 6.2. Liniami przerywanymi zaznaczono zawodników, których średnia unormowana wartość z 6 cech była dodatnia.

Obliczenie średnich wartości unormowanych dla poszczególnych osób oparte było na założeniu jednakowej ważności poszczególnych cech, tzn. waga każdej cechy była równa 1. Należy jednak pamiętać, że np. dla piłkarzy ręcznych ważniejsza może być moc maksymalna,

a dla zapaśników – czas utrzymania mocy maksymalnej i cechy te mogą mieć różne wagi, zależnie od okoliczności. Suma wag poszczególnych cech powinna być równa liczbie cech (w powyższym przykładzie 6). Jeżeli którejs cesze przypisze się (arbitralnie) wagę np. 1.5, to jednej lub kilku innym cechom należy zmniejszyć wagi, aby spełnić powyższy warunek. Następnie, wartości danej cechy należy mnożyć przez jej wagę, jak pokazano w poniższym przykładzie.

nr	$M_{sr}$	$M_{max}$	-WSM	$-T_{uz}$	$T_{utr}$	-LA	śr.
waga	$\times 0.7$	$\times 1.6$	$\times 0.7$	$\times 1$	$\times 1$	$\times 1$	
1	0.46	0.37	0.48	0.19	0.34	0.37	0.37
2	1.27	2.53	-0.22	-0.24	1.56	-1.39	0.58
3	-1.16	-2.48	-0.11	0.64	-1.91	-0.33	-0.89
4	0.19	1.86	-1.41	-0.57	-0.04	1.01	0.17
5	-0.34	-0.70	0.30	1.36	-0.32	0.8	0.18
6	-0.08	-1.02	1.02	0.62	0.75	1.01	0.38
7	-0.34	-0.53	-0.08	-1.99	-0.37	-1.46	-0.79

**Tab. 6.3** Dane z tabeli 6.2 mnożone przez arbitralnie ustalone wartości wag

Przyjęto, że dla danej dyscypliny najważniejszym parametrem mierzonym w teście Wingate jest moc maksymalna, której przypisano wagę 1.6. Za najmniej ważne uznano moc średnią i wskaźnik spadku mocy, którym przypisano wagi 0.7. Wartości pozostałych cech nie uległy zmianie, tzn. ich wagi są równe 1; tym samym suma wag jest równa liczbie cech, czyli 6.

Porównując średnie wartości dla poszczególnych zawodników z tabel 6.2 (jednakowe wagi) i 6.3 (różne wagi) można zauważyć, że w tym wypadku wprowadzenie wag nie zmieniło kolejności wyników. Również indywidualne profile nie uległy zasadniczym zmianom. Problem wag dla poszczególnych cech został tu przedstawiony raczej pobieżnie, aby dać pojęcie o tym, jak posługiwać się wagami.

### Wielocechowa analiza wariancji

Jeżeli profile wielocechowe sporządzi się nie dla jednej grupy, jak pokazano wyżej, lecz dla dwu lub więcej grup, wówczas można przeprowadzić porównania za pomocą analizy wariancji (patrz rozdział 4) ze względu na kilka cech jednocześnie. Wykorzystuje się w tym celu omówioną uprzednio zasadę addytywności, a więc algebraicznego sumowania wartości zmiennych unormowanych. Na początek należy zatem przeprowadzić we wszystkich grupach normowanie na tę samą dla danej cechy wartość średniej i odchylenia standardowego (patrz rozdział 3). Następnie należy sporządzić profile wielocechowe dla średnich wartości z poszczególnych grup, aby wyeliminować z dalszej analizy te cechy, które nie

różnicują grup. Unormowane wartości pozostałych cech zsumować algebraicznie (lub obliczyć średnie) dla poszczególnych wierszy (osób) i na tych wartościach przeprowadzić analizę wariancji.

W tabeli 6.3 przedstawiono dane dotyczące 4 wskaźników spójności grupowej w dwóch zespołach; w ostatnich dwóch wierszach każdego zespołu podano średnie i odchylenia standardowe. Wszystkie dane unormowano na średnie i odchylenia standardowe dla zespołu A (tab. 6.4), a następnie ze średnich unormowanych sporządzono profile wielocechowe (ryc. 6.b) z zaznaczonymi błędami standardowymi (SE). Ponieważ dane były unormowane na wartości z zespołu A, profil dla tego zespołu jest linią prostą na poziomie zera.

Jak widać z tabeli 6.3, zespoły wykazują różnice w zmiennych ATGS, ATGT i GIT, chociaż żadna różnica nie jest znamienna (wartości funkcji  $t$  Studenta podane są w ostatnim wierszu tabeli), natomiast średnie wartości GIS są praktycznie identyczne.

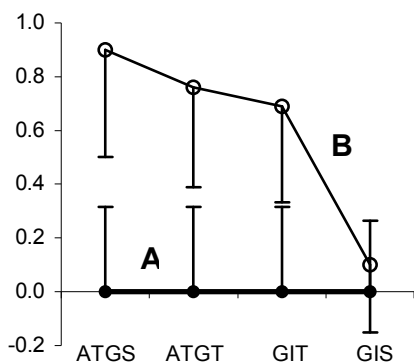
Zespół A (n = 10)				Zespół B (n = 9)			
ATGS	ATGT	GIT	GIS	ATGS	ATGT	GIT	GIS
24	22	28	21	39	15	33	18
30	24	30	19	43	27	39	20
25	13	21	9	33	27	30	10
36	25	38	24	27	27	32	21
31	19	32	10	33	27	38	20
39	24	32	11	34	26	26	10
28	21	31	18	41	27	30	15
31	25	30	20	28	18	32	22
35	16	32	17	39	26	39	24
31	24	29	24				
31.0±4.7	21.3±4.1	30.3±4.2	17.3±5.5	35.2±5.6	24.4±4.6	33.2±4.5	17.8±5.1
wartości funkcji $t$				1.79	1.56	1.45	0.09

**Tab. 6.3** Wartości parametrów spójności grupowej w dwóch zespołach siatkówki kobiet. Podane wartości funkcji  $t$  odnoszą się do różnic między średnimi obu zespołów.

W tabeli 6.4 pokazano unormowane wartości ATGS, ATGT i GIT oraz indywidualne średnie tych zmiennych. Cechę GIS pominięto, ponieważ nie różnicuje ona badanych zespołów. Średnie grupowe dla średnich z 3 cech wynoszą  $0 \pm 0.84$  i  $0.78 \pm 0.76$ , różnica między średnimi wynosi zatem 0.78, a wartość  $t$  obliczona jak podano na s. 26 jest równa 2.11. Zsumowanie trzech unormowanych zmiennych pozwoliło więc wykazać znamiennej różnicę w spójności dwóch zespołów ( $P < 0.05$ ), mimo że żadna z tych zmiennych z osobna nie różnicowała zespołów znamienne.

ATGS	ATGT	GIT	średnia	ATGS	ATGT	GIT	średnia
Zespół A				Zespół B			
-1.48	0.17	-0.54	-0.62	1.70	-1.53	0.64	0.27
-0.21	0.66	-0.07	0.12	2.55	1.39	2.05	1.99
-1.27	-2.02	-2.19	-1.83	0.42	1.39	-0.07	0.58
1.06	0.90	1.81	1.26	-0.85	1.39	0.40	0.31
0.00	-0.56	0.40	-0.05	0.42	1.39	1.81	1.21
1.70	0.66	0.40	0.92	0.64	1.14	-1.01	0.26
-0.64	-0.07	0.16	-0.18	2.12	1.39	-0.07	1.15
0.00	0.90	-0.07	0.28	-0.64	-0.80	0.40	-0.35
0.85	-1.29	0.40	-0.01	1.70	1.14	2.05	1.63
0.00	0.66	-0.31	0.12				
0.00±1.00	0.00±1.00	0.00±1.00	0.00±0.84	0.90±1.19	0.76±1.12	0.69±1.07	0.78±0.76

**Tab. 6.4** Wartości ATGS, ATGT i GIT z tabeli 6.3 unormowane na średnie i odchylenia standardowe otrzymane dla **zespołu A**. Średnie wartości z błędami standardowymi pokazano na rycinie 6.b w postaci profilów.



**Ryc. 6.b** Profil czterocechowy zespołu B dla danych unormowanych na średnie i odchylenia standardowe zespołu A. Profil zespołu A przedstawia pozioma linia leżąca na osi. Pionowe kreski oznaczają błędy standardowe (SE).

Należy tu zaznaczyć, że zastosowanie wieloczechowej analizy wariancji do takich danych, jak przedstawione w podrozdziale „Profile wieloczechowe”, może wymagać zmiany znaku dla wartości pewnych zmiennych, aby ujednolicić oczekiwany kierunek zmian. W przeciwnym razie efekty będą się wzajemnie znosić.

### Korelacje i regresje wielu zmiennych

W poprzednim rozdziale omówiono zależność jednej cechy (y) od drugiej (x) wyrażoną regresją. Cechę y można także rozpatrywać jako zależną od więcej niż jednej cechy i opisać ją *regresją wielokrotną*. Podobnie jak w regresji dwu zmiennych, warunkiem jest liniowość



zależności zmiennej  $y$  od zmiennych zależnych  $x_i$ . To samo dotyczy *korelacji wielokrotnej* – opisuje ona liniową zależność danej zmiennej  $y$  od zmiennych zależnych  $x_i$ , której miarą jest kwadrat współczynnika korelacji ( $R^2$ ; dla zależności wielocechowych oznacza się go dużą literą). Wprawdzie pierwiastek tej wartości jest technicznie równoważny zwykłemu współczynnikowi korelacji, ale nie ma on określonego znaku (plus lub minus), bowiem zależność  $y$  od  $x_1$  może być dodatnia, a od  $x_2$  – ujemna itd.

Obliczenia związane z zależnościami wielokrotnymi są uciążliwe a ich ilość szybko wzrasta z liczbą cech. Dlatego, jeśli cech jest więcej niż cztery, nie opłaca się wykonywanie rachunków nawet z pomocą Excela i lepiej użyć specjalnych programów obliczeniowych. Poniżej podano przykład korelacji wielokrotnej i regresji dla trzech cech.

### *Korelacja i regresja trzech zmiennych*

W tabeli 6.5 przedstawiono pomiary wysokości ciała, masy ciała i zawartości tłuszczu w ciele (w kg) u 22 dziewcząt w celu określenia zależności zawartości tłuszczu od wysokości i masy ciała. Zależność ta będzie przedstawiona w postaci kwadratu współczynnika korelacji ( $R^2$ ) oraz równania regresji.

h	m	f	h	m	f
117.5	21.0	5.9	151.5	48.4	10.1
115.0	23.5	6.2	141.0	35.7	5.2
126.0	25.4	6.8	139.5	49.0	11.4
124.0	25.3	7.2	143.0	36.6	5.9
125.0	30.5	8.4	135.0	30.4	6.2
120.5	20.8	3.2	141.0	34.1	6.2
135.5	36.8	10.4	138.5	29.0	4.9
132.5	30.7	7.5	128.0	32.2	7.5
141.0	31.0	5.5	147.5	39.7	10.1
140.0	37.5	8.1	147.0	41.4	7.2
135.5	37.1	10.4	144.5	43.8	8.8

**Tab. 6.5** Wysokość ciała (h), masa ciała (m) i zawartość tłuszczu w ciele w kg (f) u 22 dziewcząt.

Poniżej zestawiono obliczone wielkości niezbędne do wyznaczenia korelacji i równania regresji.

$n = 22$	$S_{hh} = 2166.45$	$B = (-0.165 \cdot 168.96) + (0.368 \cdot 251.96) = 64.84$
$\Sigma h = 2969$	$S_{mm} = 1328.51$	$R^2 = 64.84/93.65 = 0.692$
$\Sigma m = 739.9$	$S_{ff} = 93.65$	$D = 826\,724.49$
$\Sigma f = 163.1$	$S_{hm} = 1432.28$	$b_1 = -0.165$
$\Sigma h^2 = 402846.5$	$S_{th} = 168.96$	$b_2 = 0.368$
$\Sigma m^2 = 26212.69$	$S_{fm} = 251.96$	$b_0 = 17.31$
$\Sigma f^2 = 1302.81$	$r_{hm} = 0.844$	$S_{RR} = S_{ff} - B = 28.81$
$\Sigma h \cdot m = 101285.2$	$r_{hf} = 0.375$	$s_R = \sqrt{28.81/19} = 1.23$
$\Sigma h \cdot f = 22180.05$	$r_{mf} = 0.714$	$f = 17.31 - 0.165 \cdot h + 0.368 \cdot m \pm 1.23$
$\Sigma m \cdot f = 5737.31$		

Powyżej zestawiono podstawowe wielkości potrzebne do uzyskania żądanych parametrów zgodnie ze sposobem obliczeń na s. 43 oraz z wzorami podanymi na sąsiedniej stronie.

W poprzednim rozdziale wykres zależności  $y$  od  $x$  wymagał naniesienia pojedynczych punktów na dwie osi. W wypadku trzech zmiennych wykres musiałby być umieszczony w przestrzeni, bowiem 3 zmienne wymagają trzech osi:  $Y$ ,  $X_1$  i  $X_2$ . Innymi słowy, każdy punkt jest określony trzema wielkościami (współrzędnymi). Aby móc poprawnie stosować odpowiednie rachunki, należy najpierw sprawdzić rozkłady zmiennych: stosując procedury opisane na s. 14 można się łatwo przekonać, że rozkład żadnej zmiennej nie odbiega znacząco od normalnego. Nie jest zatem wymagana żadna transformacja.

W panelu na poprzedniej stronie zamieszczono sumy i sumy kwadratów trzech zmiennych oraz sumy ich iloczynów:  $f \cdot h$ ,  $f \cdot m$  oraz  $h \cdot m$ , a także odpowiadające im zredukowane sumy kwadratów i iloczynów. Z tych z kolei obliczono współczynniki korelacji ( $r$ ). Te trzy korelacje proste (dwucechowe) odpowiadają rzutom punktów w przestrzeni na trzy płaszczyzny wyznaczone osiami:  $f/h$ ,  $f/m$  i  $h/m$ .

Równanie regresji dla omawianych trzech zmiennych ma postać:

$$f = b_0 + b_1 \cdot h + b_2 \cdot m \quad [6-1]$$

Współczynniki równania w powyższym wzorze wyznacza się następująco:

$D = S_{hh} \cdot S_{mm} - S_{hm}^2$	$b_1 = (S_{mm} \cdot S_{fh} - S_{hm} \cdot S_{fm})/D$	$b_2 = (S_{hh} \cdot S_{fm} - S_{hm} \cdot S_{fh})/D$
$b_0 = (\Sigma f - b_1 \cdot \Sigma h - b_2 \cdot \Sigma m)/n$		[6-2]

Przypomnijmy (patrz s. 43), skąd pochodzi kwadrat współczynnika korelacji (prostej):

$$B = S_{xy}^2/S_{xx} = S_{xy} \cdot (S_{xy}/S_{xx}) = \mathbf{b} \cdot \mathbf{S}_{xy}, \text{ zaś } r^2 = S_{xy}^2/S_{xx} \cdot S_{yy}, \text{ czyli } r^2 = B/S_{yy}.$$

Wartość  $B$  ('regresja' jako wielkość matematyczna, a nie jako forma zależności) w korelacji wielu zmiennych jest określona wzorem 6-3.

$$B = b_1 \cdot S_{11} + b_2 \cdot S_{22} \quad [6-3]$$

Dla omawianych trzech zmiennych mamy zatem:

$$B = (-0.165 \cdot 168.96) + (0.368 \cdot 251.96) = 64.84,$$

zaś kwadrat współczynnika **korelacji wielokrotnej** trzech zmiennych wynosi

$$R^2 = \frac{B}{S_{yy}} \quad [6-4]$$

Analogicznie jak w regresji dwu zmiennych, różnica między 'wariancją całkowitą' ( $S_{ff}$ ) a 'regresją' ( $B$ ) podzielona przez liczbę stopni swobody ( $df$ ) jest wariancją błędu w regresji ( $s_R^2$ ). Pierwiastek z tej ostatniej jest błędem w regresji, który wynosi tu 1.23. Należy zauważyć, że liczba stopni swobody jest równa  $n - k$ , gdzie  $k$  jest liczbą zmiennych. W tym wypadku zatem  $df = 19$ .

Aby określić znamienność regresji zawartości tłuszczu na wysokość i masę ciała jednocześnie, obliczamy wartość  $F = B/s_R^2$ . W tym wypadku wynosi ona 42.8 dla  $df = 2, 19$  ('regresja' ma dwa stopnie swobody, tyle ile zmiennych niezależnych). Wartość tablicowa przy  $P = 0.01$  wynosi 5.93, regresja jest zatem wysoce znamienna.

Należy jeszcze stwierdzić, czy obie zmienne niezależne (wysokość i masa ciała) mają znamienny wpływ na zawartość tłuszczu, czy też tylko np. masa ciała jest istotnym czynnikiem. Można to rozwiązać na dwa równoważne sposoby: obliczając wartości testu  $t$  dla obu współczynników regresji lub obliczając *korelacje cząstkowe* (patrz niżej).

Wartości  $t$  oblicza się następująco:  $t = b/s_b$ , gdzie  $b$  jest danym współczynnikiem regresji, a  $s_b$  jego błędem;  $s_{b1} = s_R \cdot \sqrt{S_{mm}}/D$ ,  $s_{b2} = s_R \cdot \sqrt{S_{hh}}/D$ .

Podstawiając obliczone wartości  $s_R$ ,  $D$ ,  $S_{mm}$  i  $S_{hh}$  zamieszczone w panelu obliczeń (s. 77) otrzymamy:

$$s_{b1} = 1.23 \cdot \sqrt{1328.51/826724.49} = 0.050 \quad t_{b1} = -0.165/0.050 = -3.30 \quad P < 0.01$$

$$s_{b2} = 1.23 \cdot \sqrt{2166.45/826724.49} = 0.063 \quad t_{b2} = 0.368/0.063 = 5.84 \quad P < 0.001$$

Jak widać, oba współczynniki regresji są wysoce znamienne, zatem i wysokość i masa ciała mają znamienny wpływ na zawartość tłuszczu w ciele. Wartość  $R^2$  wynosi 0.692, co oznacza, że 69.2% całkowitej wariancji zawartości tłuszczu daje się wyjaśnić wysokością i masą ciała, a 30.8% - innymi, niekontrolowanymi czynnikami.

Obliczony błąd w regresji ( $s_R$ ) wyniósł 1.23 kg; jest to błąd (odchylenie standardowe) dowolnego, pojedynczego pomiaru od linii regresji, a więc od wartości wyznaczonej rachunkowo z wysokości i masy ciała. Błąd ten można wyrazić względem średniej zawartości tłuszczu:  $1.23/7.41 = 0.166$  lub 16.6%. Jak na oszacowanie wyłącznie na podstawie wysokości i masy ciała nie jest to duży błąd.

Dla sprawdzenia równania regresji policzmy oczekiwaną zawartość tłuszczu dla pierwszej (1) i ostatniej (22) dziewczynki, podstawiając do równania pomiary wysokości i masy ciała:

$$(1) f = 17.31 - 0.165 \cdot 117.5 + 0.368 \cdot 21.0 = 5.65,$$

$$(22) f = 17.31 - 0.165 \cdot 144.5 + 0.368 \cdot 43.8 = 9.58.$$

Zmierzone zawartości tłuszczu wynosiły odpowiednio 5.9 i 8.8 kg, więc w tych wypadkach różnica między zmierzoną i oszacowaną wartością była znacznie mniejsza niż błąd w regresji.

### *Korelacje cząstkowe*

Korelacje proste mierzą tylko zależność statystyczną między dwiema cechami, nie pozwalając na wyciąganie wniosków przyczynowych, bowiem ewentualny wpływ innych cech na tę zależność pozostaje poza kontrolą. Rozpatrując trzy zmienne jednocześnie, można nie tylko obliczyć współczynniki korelacji prostej dla wszystkich kombinacji cech, ale także badać zależność jednej cechy od drugiej przy wyłączeniu wpływu trzeciej zmiennej.

Aby obliczyć współczynnik korelacji cząstkowej, należy użyć współczynników korelacji prostej dla poszczególnych par rozpatrywanych cech wg wzoru 6-5.

$$r_{1,2*3} = \frac{r_{1,2} - r_{1,3} \cdot r_{2,3}}{\sqrt{(1 - r_{1,3}^2) \cdot (1 - r_{2,3}^2)}} \quad [6-5]$$

Jeżeli oznaczymy zawartość tłuszczu (f) jako 1, wysokość ciała (h) jako 2, a masę ciała (m) jako 3, to, podstawiając otrzymane w omówionym wyżej przykładzie korelacje do wzoru 6-4, otrzymamy dwie korelacje cząstkowe: zależność między zawartością tłuszczu a wysokością ciała przy wyłączeniu wpływu masy ciała ( $r_{fm*m}$ ) oraz między zawartością tłuszczu a masą ciała przy wyłączeniu wpływu wysokości ciała ( $r_{fm*h}$ ), jak pokazano w obliczeniach poniżej.

$$r_{fh*m} = \frac{0.375 - 0.714 \cdot 0.844}{\sqrt{(1 - 0.714^2) \cdot (1 - 0.844^2)}} = -0.606 \quad r_{fm*h} = \frac{0.714 - 0.375 \cdot 0.844}{\sqrt{(1 - 0.714^2) \cdot (1 - 0.844^2)}} = 0.799$$

Pierwsza z tych korelacji jest ujemna i wynosi  $-0.606$  ( $P < 0.01$ ), co oznacza, że gdyby masa ciała była u wszystkich dziewcząt taka sama, to zawartość tłuszczu byłaby tym większa, im niższe byłyby dziewczęta. Należy zwrócić uwagę, że korelacja prosta ( $r_{fm}$ ) była dodatnia i prawie znamienna ( $0.05 < P < 0.10$ ); świadczy to o tym, że prosta korelacja jest wypadkową różnych wpływów i nie może być traktowana jako miara zależności w sensie przyczynowym. Dodatnia korelacja prosta między zawartością tłuszczu a wysokością ciała była efektem bardzo wysokiej korelacji między wysokością i masą ciała ( $r_{hm} = 0.844$ ), natomiast korelacja cząstkowa pozwoliła wyeliminować ten efekt, pozostawiając „czystą” zależność zawartości tłuszczu od wysokości lub masy ciała. Korelacje cząstkowe dają zatem podstawę do wnioskowania o przyczynach, co nie jest dopuszczalne w wypadku korelacji prostych.

Druga korelacja cząstkowa ( $r_{fm*h}$ ) pozostaje dodatnia ( $0.799$ ;  $P < 0.001$ ) i niewiele wyższa od korelacji prostej ( $r_{fm} = 0.714$ ). Oznacza to, że przy stałej wysokości ciała zawartość tłuszczu rośnie z masą ciała, przy czym wysokość ciała nie ma w istocie większego wpływu na tę zależność.

Żeby uwidocznic różnorodność zależności, jakie mogą zachodzić między korelacjami prostymi i cząstkowymi, rozpatrzmy jeszcze jeden przykład dotyczący tej samej grupy dziewcząt – zależności między czasem zwisu na ugiętych ramionach (c; próba BAH z baterii EUROFIT) a wysokością (h) i masą (m) ciała. Korelacje proste wynosiły:  $r_{ch} = 0.182$ ,  $r_{cm} = -0.157$ ,  $r_{hm} = 0.844$ , zaś korelacje cząstkowe -  $r_{ch*m} = 0.594$  ( $P < 0.01$ ),  $r_{cm*h} = -0.589$  ( $P < 0.01$ ). Obie korelacje cząstkowe były znamienne wyższe od odpowiednich korelacji prostych, co łatwo sprawdzić metodą opisaną w sekcji „Porównywanie współczynników korelacji” (s. 55). W tym wypadku obie korelacje proste dla próby BAH były bliskie zeru,

świadcząc pozornie o braku zależności tej próby od wysokości bądź masy ciała, natomiast korelacje cząstkowe ujawniły wysoce znamienne, przeciwnie skierowane zależności. I tak przy stałej wysokości ciała czas zwisu zmniejszał się z masą ciała (trudniej utrzymać większy ciężar), przy stałej masie ciała zaś zwiększał się z wysokością ciała.

Przez analogię do korelacji cząstkowych można również obliczać regresje cząstkowe, a więc ilościową zależność między zmienną zależną ( $y$ ) a jedną ze zmiennych niezależnych przy wyłączeniu wpływu drugiej zmiennej niezależnej. Temat ten nie będzie tu jednak omawiany.

### *Korelacje wielu zmiennych*

Podobnie jak w wypadku trzech zmiennych można badać zależności między większą liczbą cech. Rachunki znacznie się jednak komplikują ze wzrostem liczby zmiennych i należy korzystać wówczas z programów komputerowych.

Pierwszym krokiem powinno być zbadanie rozkładów poszczególnych cech, aby dokonać ewentualnej transformacji, bowiem błędy we współczynnikach korelacji prostych wynikające z odstępstw od rozkładu normalnego będą ulegały zwielokrotnieniu ze wzrostem liczby zmiennych. Następnie układa się tzw. macierz współczynników korelacji między wszystkimi cechami metodą „każda z każdą”. Liczba wszystkich współczynników korelacji jest równa  $n \cdot (n-1)/2$ . Przy dużej liczbie cech trudno jest jednak analizować wszystkie zależności, dlatego stosuje się grupowanie cech. Można to wykonać za pomocą tzw. analizy czynników lub poprzez arbitralne grupowanie cech w zespoły po uprzednim ich unormowaniu. Analiza czynników umożliwia wykrycie zależności trudnych do uchwycenia w inny sposób, jednak tak wydzielone czynniki mogą nastręczać trudności interpretacyjne. Dlatego bardzo ważne jest zdanie sobie sprawy z ograniczeń zautomatyzowanej analizy z jednej strony i arbitralności wyboru z drugiej i branie pod uwagę pewnego wycucia badawczego w wyborze metody i interpretacji wyników.

Podobnie jak w wypadku trzech cech, dla których oblicza się korelacje cząstkowe z wyłączeniem wpływu jednej zmiennej, można obliczać takie korelacje z wyłączeniem wpływu dwu lub więcej cech na raz. Również i w tym wypadku należy korzystać z odpowiednich programów komputerowych i wybierać tylko te cechy, które mogą mieć znaczenie. Aby uniknąć trudności interpretacyjnych, liczba cech niezależnych (lub ich zespołów) mających znamienny wpływ na zmienną zależną nie powinna jednak przekraczać kilku, przy czym znamienność danej cechy jest określona znamiennością współczynnika regresji wielokrotnej ( $\beta$ ) dla tej cechy. Odpowiednie rachunki można przeprowadzić np. za pomocą programu STATISTICA.

## 7. ANALIZA DANYCH NIEPARAMETRYCZNYCH

Cechy dyskretne, a więc liczebności, nie mają rozkładu normalnego, dlatego do ich analizy nie stosuje się na ogół tzw. testów parametrycznych (test  $t$  Studenta itp.). Do oceny liczebności (częstości, frekwencji) służy funkcja  $\chi^2$  (chi-kwadrat), która nie zależy od typu rozkładu. Może być ona stosowana wszędzie tam, gdzie coś się da policzyć. Typowym przykładem są badania ankietowe, w których zlicza się liczby odpowiedzi na dane pytanie, a następnie porównuje częstości odpowiedzi, np. w różnych grupach.

Ocena danych polega na porównaniu liczebności zaobserwowanych (empirycznych) z oczekiwanymi. Te ostatnie oblicza się zakładając, że częstości odpowiedzi w różnych grupach są od siebie niezależne, dlatego ta forma testu chi-kwadrat nazywana jest testem niezależności.

### Tablice dwu- i wielozdzielne

Dla przykładu rozpatrzmy przykład częstości współwystępowania koloru oczu i włosów w losowej próbie kobiet z Półwyspu Helskiego. Ponieważ częstości włosów jasnych i ciemnych zaobserwowane w badanej grupie były jednakowe, to zakładając niezależność, powinno się oczekiwać po 20 osób jasnookich wśród jasno- i ciemnowłosych oraz po 48 osób ciemnookich wśród jasno- i ciemnowłosych, jak pokazano w prawej tabelce poniżej. Wartość funkcji chi-kwadrat jest sumą kwadratów różnic między częstościami zaobserwowanymi ( $n$ ) i oczekiwanymi ( $f$ ) podzielonych przez częstości oczekiwane. W poniższym przykładzie  $\chi^2 = (26-20)^2/20 + (14-20)^2/20 + (42-48)^2/48 + (54-48)^2/48 = 5.10$

$$\chi^2 = \sum \frac{(n-f)^2}{f} \quad [7-1]$$

Częstości zaobserwowane			
Włosy Oczy	Jasne	Ciemne	$\Sigma$
Jasne	26	14	40
Ciemne	42	54	96
$\Sigma$	68	68	136

Częstości oczekiwane			
Włosy Oczy	Jasne	Ciemne	$\Sigma$
Jasne	20	20	40
Ciemne	48	48	96
$\Sigma$	68	68	136

**Tab. 7.1** Obserwowane i oczekiwane częstości występowania jasnych bądź ciemnych włosów i oczu w losowej próbie z populacji kobiet z Półwyspu Helskiego

W podobny sposób można obliczyć wartość  $\chi^2$  dla tablicy o dowolnym wymiarze. Liczba stopni swobody dla danej tabeli wynosi:  $df = (k - 1) \cdot (w - 1)$ , gdzie  $k$  – liczba kolumn, a  $w$  – liczba wierszy. W tak zdefiniowanej wartości funkcji  $\chi^2$  obowiązuje ograniczenie, że obserwowana liczebność nie może być w żadnej klatce mniejsza od 5.

Rachunki są znacznie uproszczone, jeżeli stosuje się obliczenia oparte na logarytmach naturalnych; mianowicie, dla każdej klatki oblicza się wartość

$$n \cdot \ln(n)$$

gdzie  $n$  jest liczebnością w danej klatce. Wartości te, do  $n = 439$ , można odczytać z tablicy 8. Należy obliczyć dwie wielkości:

$A = 26 \cdot \ln(26) + 14 \cdot \ln(14) + 42 \cdot \ln(42) + 54 \cdot \ln(54) + 136 \cdot \ln(136) = 1162.17$  (klatki zacienione; przykład:  $26 \cdot \ln(26) = 84.71$ ),

$B = 40 \cdot \ln(40) + 96 \cdot \ln(96) + 68 \cdot \ln(68) + 68 \cdot \ln(68) = 1159.59$  (klatki jasne).

Następnie oblicza się wartość  $G = 2 \cdot (A - B) = 5.16$ , która jest równoważna wartości  $\chi^2$ . Poza prostotą rachunków, w metodzie tej nie ma ograniczeń co do liczebności poszczególnych klatek – mogą nawet wystąpić wartości zerowe.

Ponieważ wartość  $\chi^2$  w powyższym przykładzie jest większa od tablicowej przy  $df = 1$  i  $P = 0.05$  ( $5.16 > 3.84$ ), odrzucamy hipotezę zerową o niezależnym rozkładzie koloru oczu i włosów i stwierdzamy, że te dwie cechy są współzależne: jasne oczy i włosy, bądź ciemne oczy i włosy występują znacznie częściej niż kombinacja jasne włosy – ciemne oczy, bądź ciemne włosy – jasne oczy.

Ponieważ badana próba była losowa, można stwierdzić, że jasne włosy występują u kobiet tak samo często jak ciemne. Gdyby jasne i ciemne oczy występowały jednakowo często (po 50%), oczekiwane liczebności wyniosłyby również po 68, podobnie jak dla włosów. Zaobserwowano natomiast 40 kobiet z jasnymi oczami i 96 z ciemnymi. Różnica w stosunku do oczekiwanych wartości (68) wynosi po 28 dla jasnych i ciemnych oczu. Hipotezę o jednakowej częstości występowania jasnych i ciemnych oczu można zatem sprawdzić następująco:  $\chi^2 = 2 \cdot 28^2 / 68 = 23.06$ ; wartość ta jest dużo większa od tablicowej 10.8 ( $df = 1$ ;  $P = 0.001$ ). Włosy ciemne występują zatem wysoce znacznie częściej niż jasne.

Wreszcie, z wartości  $\chi^2$  (lub funkcji  $G$ ) można obliczyć wartość odpowiadającą stosunkowi korelacyjnemu ( $R$ ), którego poziom znamienności odpowiada funkcji  $\chi^2$ . W podanym przykładzie  $R = \sqrt{5.16/136} = 0.195$ .

$$R = \sqrt{\frac{\chi^2}{n}}$$

[7-2]

Rozpatrzmy jeszcze nieco bardziej skomplikowany przykład: przeprowadzono dwa losowe sondaże opinii publicznej na temat popularności partii politycznych (tab. 7.2).

Partia	VI.04	I.05	$\Sigma$
P	40	46	86
R	43	28	71
S	18	20	38
T	10	14	24
$\Sigma$	111	108	219

**Tab. 7.2** Wyniki dwukrotnego sondażu popularności czterech partii politycznych (liczby respondentów)

$G = 4.35$  ( $df = 3$ )

Partia	VI.04	I.05	$\Sigma$
R	43	28	71
Inne	68	80	148
$\Sigma$	111	108	219

**Tab. 7.3** Porównanie wyników partii R z wynikami pozostałych partii łącznie

$$G = 4.13 \text{ (df} = 1; P < 0.05)$$

Obliczamy funkcję G jak w poprzednim przykładzie:

$$A = 40 \cdot \ln(40) + 46 \cdot \ln(46) + \dots + 219 \cdot \ln(219)$$

$$B = 86 \cdot \ln(86) + 71 \cdot \ln(71) + \dots + 108 \cdot \ln(108)$$

$$G = 4.35 \text{ (< } 7.81 \text{ dla } df = 3).$$

Wynik ten oznacza, że nie ma podstaw do odrzucenia hipotezy zerowej dla całej tablicy, tj. poparcie dla wszystkich partii razem nie zmieniło się. Ponieważ jednak wartość  $G = 4.35$  jest większa od 3.84 (dla  $df = 1$ ), być może dla jednej z partii poparcie uległo istotnej zmianie: największą różnicę zaobserwowano dla partii R, którą porównamy z pozostałymi partiami (tab. 7.3). Wartość G dla tej tabeli wynosi 4.13, a więc więcej niż 3.84 dla  $df = 1$ . Wartość 4.13 jest tylko nieznacznie mniejsza niż 4.35 otrzymana dla pełnej tabeli powyżej. Znaczy to, że partia R znamienne straciła na popularności, zyskały zaś po trochu pozostałe partie, ale żadna z nich znamienne, co można łatwo sprawdzić, porównując kolejno wyniki dla poszczególnych partii z pozostałymi partiami łącznie.

Podsumowując, jeżeli tabela ma dwa lub więcej stopni swobody, możliwe są trzy sytuacje.

1. Obliczona wartość G jest mniejsza niż **3,84** (wartość dla **jednego** stopnia swobody); wynik taki oznacza, że w tabeli danych brak jakichkolwiek znamienych różnicowań.
2. Obliczona wartość G jest zawarta między 3,84 a wartością tablicową dla danej liczby stopni swobody; wynik taki oznacza, że niektóre dane w tabeli **mogą być** znamienne różnicowane.
3. Obliczona wartość G jest większa lub równa wartości tablicowej dla danej liczby stopni swobody (np. dla  $df = 2$  będzie to 5,99); wynik taki oznacza, że niektóre dane w tabeli **na pewno są** znamienne różnicowane.

W sytuacjach (2) i (3) konieczna jest zatem dalsza analiza w celu wykrycia konkretnych różnicowań, nie wystarczy bowiem stwierdzenie, że dane w tabeli wykazują znamienne różnicowanie, bo nie wiadomo które – czy wszystkie ze wszystkimi? To może nie być (i najczęściej nie jest) prawdą.

#### *Nieparametryczny test różnicy między dwiema grupami (test topologiczny)*

Test ten stosuje się wówczas, gdy zmienności danych w obu grupach różnią się znacznie, a rozkłady nie są określone. Poniższy przykład dotyczy wyników próby BAH (czas zwisu na ugiętych ramionach; w sekundach) u 11-letnich dziewcząt (A) i chłopców (B). Jak widać, rozrzuty są bardzo duże i test  $t$  nie może być tu zastosowany. Aby zastosować test topologiczny, należy połączyć wyniki obu grup i uporządkować (A + B). Dane dla chłopców podkreślono.



**A:** 0, 0, 0, 0, 1, 20, 20, 26, 36, 38, 55, 57, 65, 69, 80, 90, 108, 137, 140, 145

**B:** 0, 10, 22, 39, 90, 90, 100, 112, 113, 130, 134, 158, 162, 274, 279, 281, 322

**A+B:** 0, 0, 0, 0, 0, 1, 10, 20, 20, 22, 26, 36, 38, 39, 55, 57, 65 **M** 69, 80, 90, 90, 90, 100, 108, 112, 113, 130, 134, 137, 140, 145, 158, 281, 322

	<M	>M	Σ
A	13	4	17
B	7	10	17
Σ	20	14	34

**Tab. 7.4** Zestawienie częstości wyników mniejszych (<M) i większych (>M) od środka (M) łącznego zbioru (A+B) w grupie dziewcząt (A) i chłopców (B)

Następnie dzieli się zbiór na pół (środek zbioru zaznaczony literą **M**) i zlicza, ile wartości z każdej grupy jest mniejszych od M i większych od M. Zestawienie wyników zliczeń pokazano w tabeli 7.4. Na koniec, oblicza się wartość funkcji G jak w przykładzie powyżej. Wartość ta wynosi 4.52, jest więc większa od tablicowej 3.84 (df = 1; P = 0.05). Wniosek – chłopcy osiągają znamienne wyższe wartości niż dziewczęta, ponieważ w grupie A przeważają wartości mniejsze od M, a w grupie B - większe.

#### *Korelacja rangowa (Spearmana)*

W sytuacjach takich, jak opisano powyżej dla testu topologicznego, może zająć konieczność zbadania zależności między dwiema seriami pomiarów. W tabeli 7.5 pokazano wyniki testu BAH i masy ciała piętnastu 11-letnich chłopców. Pary wyników uszeregowano najpierw według wartości BAH i przypisano im kolejne rangi, a następnie według masy ciała, również przypisując im rangi. Ponieważ w szeregu BAH występują 3 wartości zerowe, nie można im przypisać rang 1, 2 i 3, lecz jedną wspólną, będącą średnią tych trzech rang. Podobnie, występują dwie wartości 20, zatem zamiast rang 5 i 6 obie otrzymują średnią – 5.5. Następnie oblicza się różnice odpowiadających sobie rang BAH i masy ciała i podnosi je do kwadratu (kolumna  $\Delta^2$ ). Korelację Spearmana oblicza się z wzoru 7-3.

$$R_s = 1 - \frac{6 \cdot \Sigma \Delta^2}{n \cdot (n^2 - 1)} \quad [7-3]$$

Dla podanego przykładu  $R_s = 1 - 6 \cdot 860.5 / 15 \cdot 224 = -0.54$

Znamienność  $R_s$  dla  $n > 10$  można odczytać z tablicy 6 dla współczynnika korelacji liniowej, w której za liczbę stopni swobody przyjmujemy  $n$ . Dla liczebności nieprzekraczającej 10 korzystamy z tabeli 7.6.

BAH		Masa ciała		$\Delta^2$
Ranga	(s)	(kg)	Ranga	
2	0	54.2	12	100
2	0	60.1	14	144
2	0	60.9	15	169
4	10	39.7	7	9
5.5	20	36.9	3	6.25
5.5	20	38.9	6	0.25
7	26	56.5	13	36
8	57	35.2	2	36
9	65	42.6	8	1
10	69	47.1	10	0
11	80	43.1	9	4
12	90	53.4	11	1
13	108	37.6	5	64
14	137	33.2	1	169
15	162	37.1	4	121
$\Sigma$				860.5

**Tab. 7.5** Wartości testu BAH i masy ciała, odpowiadające im rangi oraz kwadraty różnic między rangami ( $\Delta^2$ )

n \ P	0.05	0.01
5	0.90	-
6	0.83	0.95
7	0.72	0.90
8	0.65	0.84
9	0.60	0.79
10	0.57	0.75

**Tab. 7.6** Wartości współczynników korelacji Spearmana dla liczebności  $n = 5 \div 10$  i poziomów znamienności  $P = 0.05$  i  $0.01$

Omówione powyżej testy nieparametryczne zostały wybrane spośród wielu podobnych ze względu na ich prostotę i szerokie stosowanie. Należy jednak zauważyć, że funkcja chi-kwadrat nie ogranicza się do zmiennych dyskretnych, lecz bywa również stosowana do zmiennych ciągłych, np. w ocenie podobieństwa profilów wielocechowych. Temat ten, jako zbyt specjalistyczny, nie będzie tu omawiany.

## 8. ZASADY TWORZENIA NORM I POSŁUGIWANIA SIĘ NIMI

Termin „norma” bywa potocznie używany w różnych znaczeniach, może zatem prowadzić do nieporozumień. Np., Słownik Języka Polskiego definiuje normę jako „*ilość, miara, granica ustalona, przewidziana jako wymagana lub obowiązująca w jakimś zakresie*”; Pawelski i Maj („Normy i kliniczna interpretacja ...”) określają normę jako „*umowną wartość prawidłową, stwierdzaną u osób 'klinicznie' zdrowych*”.

Na pojęcie normy składają się dwa elementy - aksjologiczny, postulatywny i empiryczny, opisowy. Jeden z tych elementów może przeważać, ale ich całkowite rozdzielenie było by trudne. Można przyjąć, że na ogół normy postulatywne odnoszą się do pojęć bardziej ogólnych, jak zdrowie, dobre samopoczucie, a także do określonych wymogów, koniecznych do osiągnięcia w danej dziedzinie (np. określony poziom wydolności w sporcie). Normy opisowe mają natomiast zastosowanie do pojedynczych cech, bądź ich kombinacji.

Omówione tu będą normy opisowe dla pojedynczych cech, stosowane w naukach przyrodniczych, np. w medycynie. Ponieważ w takich normach element opisowy zdecydowanie przeważa nad postulatywnym (często bowiem trudno jest jednoznacznie określić jaki „powinien być” zakres wartości danej cechy), a opis cechy wynika z jej zbadania w grupie uznanej według innych kryteriów za „prawidłową”, właściwiej jest posłużyć się terminem *wartości referencyjne* zamiast terminem *norma*.

### Wartości referencyjne

Aby daną wartość, np. wynik pomiaru jakiejś cechy, można było porównać z „normą”, tj. ocenić czy i na ile wartość ta mieści się w przedziale określonym jako „norma”, należy takim przedziałem dysponować. Ponieważ nie jest możliwe uzyskanie takiego przedziału („normy”) w sensie aksjologicznym, bo nie można *a priori* orzec, jaki „powinien być” ten przedział, należy go określić drogą empiryczną. Przede wszystkim, należy dokładnie zdefiniować klasę osób, do której ten przedział ma się odnosić, nie można bowiem stosować do dzieci „normy” wzrostu człowieka dorosłego. Wynika z tego ważny wniosek, że w zasadzie nie ma „norm” uniwersalnych, tylko cząstkowe.

Po zdefiniowaniu klasy osób, których ma dotyczyć „norma” danej cechy, należy wybrać drogą losową dostateczną liczbę osób spełniających kryteria przynależności do tej klasy, stanowiących tzw. próbę reprezentatywną (zob. s. 7). Wybrane osoby powinny ponadto spełniać kryteria „normalności” w sensie postulatywnym, a więc np. odpowiednio zdefiniowane kryterium zdrowia, przy czym cecha, dla której ma się ustalić zakres „normy” nie może być brana pod uwagę w definiowaniu kryterium zdrowia.

Poprawnie zmierzone wartości danej cechy w grupie osób zdefiniowanych jak wyżej, stanowią podstawę do określenia przedziału „wartości referencyjnych”.

## Cechy i rozkłady ich wartości

Jedno- lub wieloszczytowość rozkładu wiąże się z jego jednorodnością. Wyróżniamy dwa rodzaje jednorodności (zob. s. 11): statystyczną i stochastyczną (losową). Rozkład, który spełnia statystyczne kryteria jednorodności (jednoszczytowość, oceniana często tzw. miarą spłaszczenia), nazywamy jednorodnym statystycznie. Rozkład wartości pochodzących z losowej, reprezentatywnej i biologicznie jednorodnej populacji, jest rozkładem jednorodnym stochastycznie i musi być jednocześnie rozkładem jednorodnym statystycznie. Warunek „jednorodności biologicznej” jest najczęściej trudny do osiągnięcia, nawet mimo starannej selekcji osób spełniających ostre kryteria, dlatego najczęściej porzeczkuje się na ustaleniu jednorodności statystycznej. Należy jednak pamiętać, że chociaż jednorodność stochastyczna implikuje jednorodność statystyczną, to nie odwrotnie i rozkład jednorodny statystycznie nie musi (i często nie jest) jednorodny stochastycznie.

Należy zwrócić uwagę, że testy statystyczne służące do oceny jednorodności rozkładu (np. Kołmogorowa-Smirnowa, Shapiro-Wilka, chi-kwadrat itp.) są przydatne w analizie statystycznej, np. w analizie wariancji, regresji itp., są natomiast zbyt mało czułe dla potrzeb konstrukcji przedziałów referencyjnych. Właściwsze jest tu posługiwanie się badaniem zgodności dystrybuanty otrzymanej z teoretyczną, które pozwala wykryć niejednorodność tam, gdzie zawodzą wspomniane wyżej testy (por. s. 11 – 16).

Jeżeli rozkład wartości cechy jest jednorodny (statystycznie), należy sprawdzić, czy jest on wystarczająco zgodny z rozkładem normalnym, aby można było poprawnie zastosować konwencjonalne miary opisu (średnia, odchylenie standardowe). Zwykle jednak, zwłaszcza przy dużych próbach, odchylenia od normalności są zbyt duże, aby można było stosować miary standardowe.

Niekiedy udaje się doprowadzenie rozkładu skośnego do postaci symetrycznej, zbieżnej z rozkładem normalnym, poprzez transformację danych. Obowiązują tu pewne reguły, zależne od rodzaju cechy, wspomniane na stronach 20 – 23.

Osobnym zagadnieniem jest transformacja wartości będących różnicami (np. przyrostów) – nie ma tu jednolitych reguł i można napotkać znaczne trudności. Na ogół właściwsze jest stosowanie ilorazów (np. wartość „po” dzielona przez wartość „przed”) z jednoczesnym logarytmowaniem.

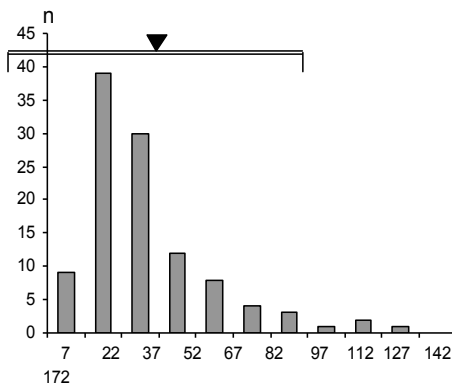
## Ustalenie przedziału referencyjnego

Poprawnie ustalony przedział referencyjny zawiera określoną frakcję, np. 95%, populacji pomiarów danej cechy, przy czym frakcje wartości mniejszych od dolnej i większych od górnej granicy przedziału powinny być w przybliżeniu jednakowe. Warunek ten będzie osiągnięty, gdy rozkład wartości jest jednorodny i symetryczny, zgodny z rozkładem normalnym. Do takich wartości można stosować statystyczne miary opisu – średnią arytmetyczną i odchylenie standardowe. Jeżeli natomiast te miary statystyczne zastosuje się do

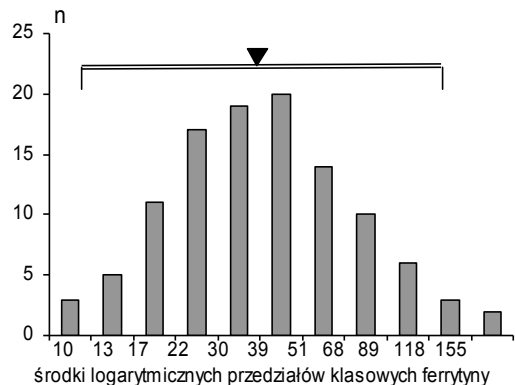
rozkładu np. skośnego, otrzyma się niedobór wartości poniżej dolnej granicy wyznaczonego przedziału, a nadmiar powyżej górnej granicy. Poprawnie zastosowana transformacja zapobiega temu.

Wypada w tym miejscu przypomnieć pojęcie błędu I i II rodzaju, przedstawione w uproszczeniu na ryc. 4.a. W niemal każdym postępowaniu diagnostycznym obowiązuje zasada minimalizacji ryzyka. W rozpoznaniu medycznym należy minimalizować ryzyko uznania osoby chorej za zdrową (błąd II rodzaju), kosztem zwiększenia liczby osób zdrowych uznanych za prawdopodobnie chore (błąd I rodzaju), bowiem tę grupę „podejrzanych o chorobę” można poddać dalszym badaniom. Wpływ rozkładu na przedział referencyjny widać na ryc. 8.a. i 8.b. (powtórzenie ryc. 2.b. na s. 11), na których pokazano rozkład wartości ferrytyny i przedział  $\bar{x} \pm 2SD$  obliczony z wartości surowych i po przekształceniu logarytmicznym.

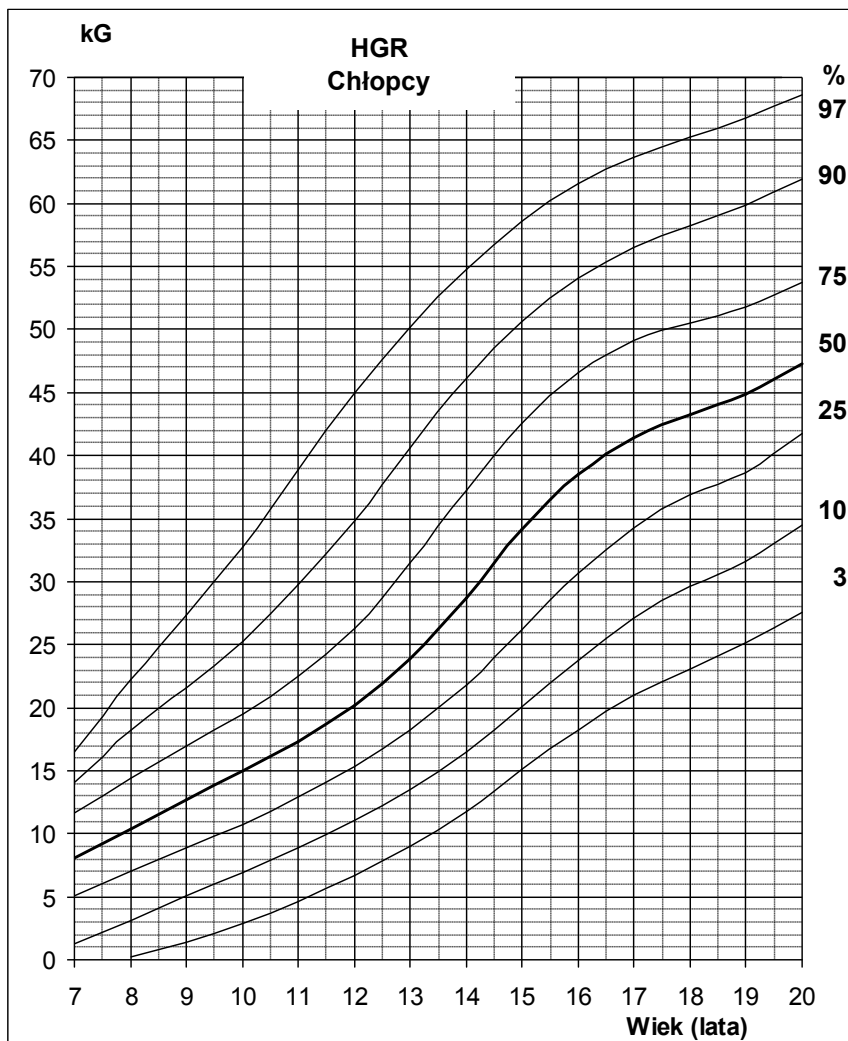
Należy zaznaczyć, że podane tu wartości stężeń ferrytyny nie stanowią przedziału referencyjnego, a są jedynie ilustracją wpływu asymetrii rozkładu na ustalanie takiego przedziału. W powyższym przykładzie rozkład obejmował dane uzyskane zarówno od osób zdrowych, jak i ze stwierdzonym niedoborem zasobów żelaza.



**Ryc. 8.a.** Rozkład stężeń ferrytyny. Podwójną kreską zaznaczono zakres średniej  $\pm$  dwa odchylenia standardowe (od  $-17$  do  $101$ )



**Ryc. 8.b.** Logarytmiczny rozkład stężeń ferrytyny. Podwójną kreską zaznaczono zakres średniej  $\pm$  dwa odchylenia standardowe (od  $10.3$  do  $116$ )



**Ryc. 8.c.** Przedziały centylowe dla siły ręki chłopców względem wieku kalendarzowego [wg R.Stupnicki, R.Przewęda, K.Milde (2002) Centylowe siatki sprawności fizycznej polskiej młodzieży wg testów EUROFIT. Wyd. AWF, Warszawa]

Jeżeli nie jest możliwe ustalenie przedziału referencyjnego za pomocą średniej i odchylenia standardowego, np. z powodu niejednorodności rozkładu, należy przedział ten wyznaczyć graficznie, z dystrybuanty wykreślonej na tzw. papierze probitowym (ryc. 2.k). Można wówczas wyznaczyć również standardowe przedziały centylowe, zawierające 3, 10, 25, 50, 75, 90 i 97% populacji (ryc. 8.c). Linia wyznaczająca np. 10. centyl oznacza, że poniżej tej linii mieszczą się wartości 10% populacji. Jako przykład pokazano na wykresie punkt odpowiadający chłopcu 11.5-letniemu, którego siła ręki wynosi 13 kG. Punkt ten znajduje się

między 90. i 97. centylem (czyli leży w tzw. kanale centylowym 90 – 97%), a więc chłopiec ten reprezentuje „górne” 10% populacji. Inaczej mówiąc, ponad 90% populacji w tym wieku osiąga gorsze wyniki.

Należy zwrócić uwagę, że szerokości kanałów centylowych mogą się zmieniać nie tylko z wiekiem, ale mogą być różne dla danego wieku. Jest to wynikiem wyznaczenia siatki centylowej z dystrybuanty rzeczywistego rozkładu, a nie ze średniej i odchylenia standardowego.

### Posługiwanie się przedziałami referencyjnymi

Żeby móc poprawnie posługiwać się przedziałami referencyjnymi trzeba wiedzieć, jaką frakcję wartości spotykanej w „prawidłowej” populacji zawiera dany przedział. Większość tzw. norm obejmuje zakres średniej  $\pm 2$  odchylenia standardowe, bądź przedział 95-procentowy, niekiedy jednak przedział ten jest mniejszy i obejmuje np. 80%. Niestety, podręczniki, np. diagnostyki laboratoryjnej, tabele prawidłowych wartości klinicznych itp., najczęściej nie podają ani tzw. poziomu ufności ani informacji o symetrii przedziału, co pozwala jedynie na wykorzystanie takich danych w sposób bardzo powierzchowny i nieprecyzyjny, polegający na stwierdzeniu, że dana wartość „mieści się w normie”. Omówiony poniżej sposób posługiwania się przedziałami referencyjnymi zakłada znajomość średniej i odchylenia standardowego danej cechy oraz symetrię rozkładu.

Poprawnym i precyzyjnym sposobem jest wyrażenie danego wyniku w postaci unormowanej (zob. s. 24) na średnią i odchylenie standardowe dla danego przedziału referencyjnego („normy”). Granica przedziału referencyjnego nie jest bowiem magiczną wartością, bezwzględnie różnicującą wartości prawidłowe od patologicznych, a tylko umowną, określoną pewnym prawdopodobieństwem „prawidłowości”. Wartość  $t$ , czyli wynik unormowany, wyrażony w jednostkach odchylenia standardowego danej cechy, może być zatem oceniony za pomocą tablicy funkcji  $t$  Studenta. Ma to szczególne znaczenie wówczas, gdy dysponuje się pomiarami kilku różnych cech potrzebnych do rozpoznania: mimo że każdy wynik oddzielnie mieści się w swoim przedziale referencyjnym, rozpatrywane łącznie mogą wyraźnie wskazywać na patologię. Należy tu wyróżnić dwie sytuacje: gdy pomiary kilku cech służą do (a) potwierdzenia lub wykluczenia określonego rozpoznania, bądź (b) wyboru najbardziej prawdopodobnego z kilku możliwych rozpoznań.

Dla określonego rozpoznania, rozpatrywane odchylenia od „normy” kilku cech są tylko mniejsze lub tylko większe od danego przedziału referencyjnego, zaś odchylenia w drugą stronę są ignorowane. Oblicza się wartość funkcji chi-kwadrat, będącej sumą kwadratów wartości  $t$  dla każdej cechy (patrz wyżej). Jeżeli wynik jest większy od wartości tablicowej przy liczbie stopni swobody równej liczbie rozpatrywanych cech i przy prawdopodobieństwie 0.1 (test jednostronny), rozpoznanie można uznać za znamienne. Może zaistnieć również odwrotna sytuacja: mimo że pomiar jednej z cech znajduje się poza przedziałem referencyjnym,

pomiary kilku cech wziętych łącznie mogą nie wskazywać na patologię. Ilustrację takich przypadków stanowią poniższe przykłady (tab. 8.1).

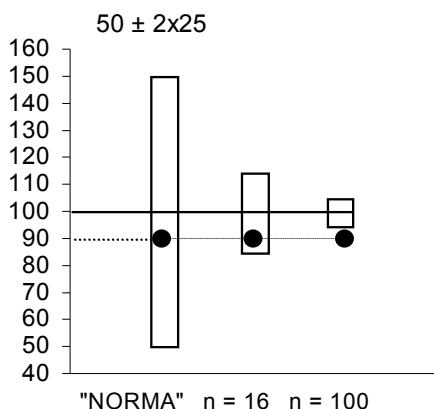
Jeżeli jest kilka możliwych rozpoznań, można rozpatrywać je po kolei jak pokazano wyżej, albo zastosować tzw. analizę profilową, również opartą na funkcji chi-kwadrat, a polegającą na określeniu stopnia podobieństwa indywidualnego profilu wielocechowego do jednego z profilów „wzorcowych”. Metoda ta, znacznie bardziej skomplikowana, nie będzie tu bliżej omawiana.

Parametr	„norma” (zakres 95%)	średnia $\pm$ SD	Pacjent A		Pacjent B	
			$x$	$z$	$x$	$z$
Ferrytyna #	>18.4	1.568 $\pm$ 0.152	15.8	2.42	19.5	1.84
Fe <sup>++</sup>	60-120	90 $\pm$ 15	85	0.33	67	1.53
CZWŻ #	200-500	2.500 $\pm$ 0.100	330	0.20	468	1.70
			$\chi^2$	6.01	$\chi^2$	8.61*

**Tab. 8.1** Indywidualne wartości surowe ( $x$ ) i unormowane ( $z$ ) ferrytyny, żelaza (Fe<sup>++</sup>) i całkowitej zdolności wiązania żelaza (CZWŻ) dwóch pacjentów (A i B) oraz obliczone na tej podstawie wartości  $\chi^2$

# Norma obliczona ze średniej i SD dla wartości logarytmicznych; \* P<0.05

Należy podkreślić, że powyższe rozważania odnoszą się tylko do pomiarów dotyczących jednej osoby, a nie grupy osób. Zdecydowanie niewłaściwe jest często spotykane stwierdzenie, że „średnia wartość w danej grupie mieściła się w przedziale normy”. Przedział referencyjny odnosi się bowiem wyłącznie do wartości indywidualnych, natomiast „norma” dla średniej jest określona przedziałem ufności dla błędu standardowego (SE) przy danym prawdopodobieństwie i liczebności grupy. Przykłady przedziałów ufności dla średniej przy różnych liczebnościach pokazano na ryc. 8.d.



**Ryc. 8.d.** Zakres „normy” dla indywidualnych wartości, przedziały ufności dla prób o liczebności  $n = 16$  i  $100$  oraz wartość średnia stwierdzona w badaniach. Jeżeli liczebność badanej próby wynosiła  $100$ , to średnia ( $90$ ) nie mieści się w przedziale  $100 \pm 2 \cdot 25/\sqrt{100}$ .



## 9. KOMPUTER – POMOŻE, CZY ZASTĄPI?

Komputery wydatnie ułatwiły rachunek statystyczny. Złożyły się na to dwie przyczyny: niewyobrażalna jeszcze w latach siedemdziesiątych i osiemdziesiątych szybkość przetwarzania danych oraz mnogość łatwo dostępnych programów komputerowych. O ile jednak ogromna szybkość obliczeń niewątpliwie ułatwia pracę, o tyle posługiwanie się programami statystycznymi, np. STATISTICA, SPSS itp., może stworzyć złudzenie, że oto mamy narzędzie, które za nas zrobi, co trzeba. Że nie jest już konieczna znajomość zasad rachunku statystycznego, że wystarczy wprowadzić dane do komputera, a reszta „zrobi się sama”. Moje wieloletnie doświadczenie w tej dziedzinie wskazuje, że wyjąwszy pewne wystandaryzowane sytuacje, jak np. kontrola jakości produkcji, pewne zagadnienia doświadczalnictwa rolniczego, wiele metod laboratoryjnych, niektóre typy prób laboratoryjnych i klinicznych w kontroli leków itp., większość prac badawczych wymaga drobiazgowego zbadania warunków doświadczenia, rozkładów zmiennych, dopasowania obliczeń do schematu doświadczenia itp. Krótko mówiąc, żaden komputer, przynajmniej jak dotąd i w granicach dostępności, nie zastąpi myślenia, wyobraźni i znajomości tematu. Zwłaszcza ta ostatnia jest niezbędna dla poprawnego zanalizowania danych i wyciągnięcia wniosków. Ci, którzy korzystali z pomocy wybitnych matematyków-statystyków i nie byli zadowoleni z wyniku na skutek trudności w porozumieniu się co do przedmiotu badań, np. przyrodniczych, łatwo zrozumieją potencjalne trudności w porozumieniu się z komputerem.

Zanim jeszcze przystąpi się do właściwej analizy danych, zebrane wyniki powinny być wprowadzone do komputera w postaci tak ułożonej tabeli, aby można było przeprowadzić odpowiednie obliczenia. Dalej zostaną omówione proste sposoby wprowadzania i przetwarzania danych za pomocą arkusza kalkulacyjnego EXCEL. Program ten pozwala ponadto na łatwą wymianę danych z innymi programami windowsowymi i pakietami statystycznymi.

### Wprowadzanie i formatowanie danych

Dane do arkusza Excela wprowadza się ręcznie lub z innego pliku, np. z tabeli Worda. Dane poszczególnych osób lub innych obiektów wprowadza się zwykle wierszami. Oznaczenia kolumn (zmiennych) lepiej jest umieszczać nie w pierwszym wierszu, a na samym końcu, zostawiając 3 – 4 puste wiersze przed nazwami (ryc. 9.a). Te puste wiersze służą do wstawienia obliczanych parametrów, np. sum, średnich, odchyłeń standardowych.

	A	B	C
1	12	5.4	8.1
2	14	4.9	8.8
3	15	6.2	7.7
4	14	6	9.1
5			
6			
7			
8	wiek	czas A	czas B

**Ryc. 9.a** Fragment arkusza EXCEL z danymi i nazwami zmiennych. w zaznaczonej komórce (A7) będzie wstawiona funkcja statystyczna. Zwrócić uwagę na czarną kropkę w prawym dolnym rogu komórki – jest to tzw. uchwyt.

Wpisywane lub przenoszone z innego programu liczby są wyrównywane do prawej strony (chyba że dano inne polecenie). Wyrównanie do lewej oznacza, że dane nie są traktowane jak liczby, tylko jak słowa i może to wynikać np. z użycia kropki zamiast przecinka lub odwrotnie. Można to sprawdzić wykonując np. automatyczne sumowanie (przycisk  $\Sigma$ ) i w razie potrzeby wykonać zamianę jednego znaku na drugi [Edycja – Zamień] po uprzednim zaznaczeniu całego arkusza (kliknięcie myszką w lewym górnym rogu arkusza). Jeżeli znak oddzielający miejsca dziesiętne jest poprawny, a mimo to dane nie są rozpoznawane jako liczby, można skopiować dane do nowego pliku Worda, a następnie spod Worda skopiować do Excela.

Dane nie powinny zawierać więcej cyfr znaczących, niż to konieczne (zob. s. 5). Jeżeli kolumna danych jest wynikiem obliczeń (patrz dalej), wartości mogą mieć nadmiar miejsc dziesiętnych; należy wówczas nadmiar zredukować przyciskiem  $\rightarrow .0$ .

Zmiana liczby miejsc dziesiętnych nie ma wpływu na wynik obliczeń, bo miejsca obcięte zostają tylko ukryte, natomiast zapisanie pod innym formatem lub przeniesienie pod Worda trwale usuwa obcięte cyfry.

### Funkcje statystyczne programu

Program EXCEL wykonuje obliczenia wielu funkcji statystycznych, np. średniej, odchylenia standardowego, regresji, współczynnika korelacji itp., jak również przeprowadza testy statystyczne, np. test  $t$  Studenta. Poniżej omówione są niektóre funkcje. Wszystkie funkcje statystyczne wywołuje się przyciskiem  $f_x$  na pasku narzędzi, a następnie wybiera żądaną funkcję (np. średnią). Jeżeli funkcji tej nie ma w sekcji „Ostatnio używane”, należy wybrać opcję „Statystyczne” i stamtąd wybrać żądaną funkcję.

#### *Średnia arytmetyczna, odchylenie standardowe*

Jeżeli odchylenie standardowe ma być wpisane pod średnią, należy zacząć od obliczenia odchylenia standardowego (zob. ryc. 9.a), a potem komórkę wyżej zaznaczyć pole dla średniej dlatego, że obliczenia są wykonywane nad zaznaczonym polem w górę, do pierwszego pustego pola za ciągiem liczb. Jeżeli najpierw byłyby obliczana średnia, a potem pod

nią odchylenie standardowe, to średnia zostałaby potraktowana jak jedna z wartości. W sytuacji, gdy mamy więcej kolumn danych, jak na rycinie 8.a, nie ma potrzeby wykonywania obliczeń dla każdej kolumny oddzielnie. Należy tylko zaczernić oba pola zawierające średnią i odchylenie standardowe, w razie potrzeby zredukować nadmiar miejsc dziesiętnych, a następnie naprowadzić myszkę na ‘uchwyt’ tak, aby pokazał się krzyżyk **+**, przycisnąć i przytrzymać przycisk myszki i przeciągnąć poziomo do ostatniej kolumny danych. Średnie i odchylenia standardowe zostaną automatycznie wstawione dla wszystkich kolumn.

Jeżeli obliczenia mają być wykonane tylko dla części danych w kolumnie, należy wywołać funkcję np. średniej i po ukazaniu się planszy danej funkcji wpisać żądany zakres adresów danych ręcznie (np. B3:B11) lub przeciągnąć myszką ten zakres. Jeżeli plansza funkcji zakrywa obszar danych, można ją przesunąć w inne miejsce przeciągając myszką.

### *Badanie rozkładu*

Analizowany zbiór wartości należy uporządkować od najmniejszej do największej wartości (A), a w sąsiedniej kolumnie (B) wpisać kolejne numery. W kolumnie obok (C) wpisać formułę:  $= (B1-0.5)/14$ , nacisnąć ENTER i przeciągnąć myszką w dół (14 jest liczbą danych). W kolumnie D wstawić funkcję ROZKŁAD.NORMALNY.S.ODW.(C1) i przeciągnąć myszką w dół.

	A	B	C	D
1	350	1	0.04	-1.80
2	390	2	0.11	-1.24
3	412	3	0.18	-0.92
4	421	4	0.25	-0.67
5	425	5	0.32	-0.46
6	443	6	0.39	-0.27
7	446	7	0.46	-0.09
8	451	8	0.54	0.09
9	456	9	0.61	0.27
10	475	10	0.68	0.46
11	491	11	0.75	0.67
12	502	12	0.82	0.92
13	516	13	0.89	1.24
14	522	14	0.96	1.80

W tabeli obok pokazano odpowiedni przykład pochodzący z danych w tabeli 2.4 (s. 14). Z danych zawartych w kolumnach A i D należy sporządzić wykres i ocenić jego liniowość (ryc. 2.g, s. 14).

### *Test t Studenta*

Po wybraniu funkcji „Test t” pokazuje się plansza zawierająca 4 paski poleceń. W pierwszym pasku należy wpisać adresy danych (lub zaznaczyć zakres myszką), przejść myszką lub klawiszem **Tab** do następnego paska i analogicznie wprowadzić drugą grupę danych, Potem przejść do następnego paska („Ślady”) i wpisać 1 lub 2, odpowiednio dla jedno- lub dwustronnego testu (por. s. 40). W pasku czwartym wpisuje się 1 w wypadku danych zależnych, 2 – dla grup niezależnych o podobnych odchyleniach standar-

dowych lub 3 – jeżeli odchylenia standardowe bardzo znacznie się różnią (niezalecane!). Wynik wpisany do uprzednio zaznaczonej komórki jest prawdopodobieństwem (hipotezy zerowej), a nie wartością funkcji  $t$ , nie ma więc potrzeby porównywania wyniku z wartościami tablicowymi.

Jest oczywiste, że gdy stosuje się test  $t$  dla danych zależnych (sparowanych), liczby danych w obu grupach muszą być takie same.

### Współczynnik korelacji

Po wybraniu funkcji „Współczynnik korelacji” ukazuje się plansza z dwoma paskami; do pierwszego wpisuje się adresy lub zaznacza myszką zakres wartości pierwszej zmiennej, do drugiego paska – drugiej zmiennej. Należy uważać, by zakresy obu zmiennych zawierały tyle samo danych.

### Regresja dwu zmiennych

Wprawdzie funkcja regresji (REGLINP) jest wbudowana, ale wygodniej jest otrzymać równanie regresji łącznie z wykresem. W tym celu należy umieścić w sąsiednich kolumnach wartości  $x$  i  $y$  (w tej kolejności), następnie zaznaczyć zakres obu zmiennych, kliknąć myszką zakładkę *Wstawianie* i wybrać opcję *Punktowy*. Następnie kliknąć myszką w którykolwiek punkt wykresu, wybrać *Układ* w pasku poleceń, następnie wybrać *Analiza, Linia trendu, Więcej opcji linii trendu* i wybrać typ regresji, a potem ew. *Wyświetl równanie na wykresie* i/lub *Wyświetl wartość R-kwadrat na wykresie*. Można w ten sposób otrzymać również równania logarytmiczne i potęgowe, a dla nich stosunek korelacyjny ( $R^2$ ). Nie można natomiast otrzymać automatycznie błędów w regresji i trzeba go obliczyć jak opisano niżej.

### Obliczenia niestandardowe

Bardzo często standardowe funkcje statystyczne są niewystarczające i trzeba wykonać inne obliczenia, np. sumy kwadratów odchyłeń, wartości funkcji  $G$  itp. W tym celu trzeba najpierw zaznaczyć pierwszą komórkę, do której ma być wpisana zadana wartość. Należy pamiętać, że wpisanie zadanej formuły należy poprzedzić znakiem równości (=).

	A	B	C
1	12	5,4	=A1*ln(A1)
2	14	4,9	
3	15	6,2	
4	14	6	

Mamy np. obliczyć wartości funkcji  $G$  dla danych z kolumny A i wpisać je w kolumnie C. Zaznaczamy komórkę c1 i piszemy:

=A1\*ln(A1), co oznacza: zawartość komórki A1 pomnożyć przez logarytm naturalny zawartości komórki A1. Następnie nacisnąć ENTER i w

zaznaczonej komórce otrzymamy wartość 29.82. Aby obliczyć pozostałe wartości funkcji dla kolumny A należy złapać myszką uchwyt komórki tak, aby pokazał się krzyżyk i przeciągnąć w dół (zob. s. 82). W ten sposób oblicza się elementy funkcji  $G$  jako formy chi-kwadrat.

Aby obliczyć sumę kwadratów odchyleń od dowolnej regresji ( $S_{RR}$ ), należy najpierw obliczyć równanie regresji, a następnie podobnie jak w przykładzie wyżej wpisać do wolnej komórki formułę:  $= (b1 - 5.33 - 0.353 * a1)^2$ , gdzie **B1** jest adresem pierwszej wartości  $y$ , **A1** – pierwszej wartości  $x$ , a  $^$  jest znakiem potęgowania. Po przeciągnięciu myszką otrzymujemy szereg wartości kwadratów odchyleń, które należy zsumować (zaznaczyć pod kolumną wolną komórkę i kliknąć przycisk  $\Sigma$ ).

Poniżej podano jeszcze niektóre inne formuły obliczeniowe przydatne w pracy oraz wykaz najczęściej używanych funkcji.

Logarytmowanie	=log10(adres)
Delogarytmowanie	=10^(adres)
Pierwiastkowanie	=pierwiastek(adres) lub =(adres)^0.5
Funkcja <i>arsin</i> $v\%$	=asin(adres^0.5/100)*57.3
Maksymalna wartość zaznaczonego zbioru	=max(adresy)
Minimalna wartość zaznaczonego zbioru	=min(adresy)

Nazwa funkcji	Zastosowanie
ŚREDNIA	
ODCH.STANDARDOWE	
WSP.KORELACJI	
SUMA.KWADRATÓW	W analizie wariancji i analizie regresji
SUMA.ILOCZYNÓW	W analizie regresji
TEST.T	Porównywanie średnich
ROZKŁAD.NORMALNY.S.ODW	Badanie rozkładu

Pokazane przykłady są oczywiście jedynie wskazówką, jak należy korzystać z możliwości obliczeniowych arkusza Excela. Możliwości te są wystarczające do wykonania prostszych obliczeń i analiz, natomiast w bardziej skomplikowanych sytuacjach należy korzystać z procedur omówionych w odpowiednich rozdziałach lub ze specjalistycznych programów statystycznych.

## 10. TABLICE STATYSTYCZNE

1. Wartości rozkładu normalnego
2. Wartości dystrybuanty rozkładu normalnego
3. Wartości funkcji  $t$  Studenta
4. Wartości funkcji  $F$  Snedecora
5. Wartości funkcji  $\chi^2$  (chi-kwadrat)
6. Wartości współczynnika korelacji prostej (Pearsona)
7. Wartości współczynnika korelacji wielokrotnej
8. Wartości funkcji  $G = [n \cdot \ln(n)]$
9. Wartości funkcji  $z$
10. Wartości funkcji  $\arcsin\%$

Objaśnienia do tablic 8 – 10:

**Tablica 8** W pierwszej kolumnie podane są dziesiątki, w pierwszym wierszu – jednostki. Wartość  $n \cdot \ln(n)$  dla liczebności 23 wynosi 72.12 .

**Tablica 9** Funkcja  $z$  służy do przekształcenia współczynników korelacji na wartości o rozkładzie normalnym. Ponieważ w miarę wzrostu współczynnika korelacji wartości  $z$  szybko zwiększają się, podano 4 zakresy wartości  $r$ . Do wartości  $r$  w pierwszej kolumnie należy dodać wartość odpowiedniego wiersza nagłówkowego i tak dla  $r = 0.18$   $z = 0.101$ ; dla  $r = 0.468$  (bo  $0.45 + 0.018$ )  $z = 0.361$ ; dla  $r = 0.812$  ( $0.80 + 0.012$ )  $z = 1.581$  itd. Dla pośrednich wartości  $r$ , nieujętych w tabelicy, wartości  $z$  otrzymuje się przez interpolację. Aby znaleźć wartość współczynnika korelacji odpowiadającą np.  $z = 1.010$ , znajdujemy najbliższe wartości  $z$  wynoszące 1.005 i 1.015, którym odpowiadają wartości  $r$  0.728 i 0.730. Wartości 1.710 odpowiada zatem  $r = 0.729$ . Ujemnym wartościom  $r$  odpowiadają ujemne wartości  $z$ .

**Tablica 10** Wartości w tej tabelicy służą do przekształcenia procentów na liczby o rozkładzie zbliżonym do normalnego. Procent stanowi sumę wartości w pierwszej kolumnie i w pierwszym wierszu (do 25%); wartość  $k$  dla 11.5% wynosi 19.9 . Powyżej 25% procenty podane są z dokładnością do 0.5%. Wielkość  $k$  dla wartości powyżej 50% jest równa  $90 - k$ , gdzie  $k$  jest obliczone dla wartości 100 – 50%.

**Tablica 1**  
**Wartości rozkładu normalnego**

<b><i>t</i></b>	<b>0.00</b>	<b>0.01</b>	<b>0.02</b>	<b>0.03</b>	<b>0.04</b>	<b>0.05</b>	<b>0.06</b>	<b>0.07</b>	<b>0.08</b>	<b>0.09</b>
<b>0.0</b>	0.399	0.399	0.399	0.399	0.399	0.398	0.398	0.398	0.398	0.397
<b>0.1</b>	0.397	0.397	0.396	0.396	0.395	0.395	0.394	0.393	0.393	0.392
<b>0.2</b>	0.391	0.390	0.389	0.389	0.388	0.387	0.386	0.385	0.384	0.383
<b>0.3</b>	0.381	0.380	0.379	0.378	0.377	0.375	0.374	0.373	0.371	0.370
<b>0.4</b>	0.368	0.367	0.365	0.364	0.362	0.361	0.359	0.357	0.355	0.354
<b>0.5</b>	0.352	0.350	0.348	0.347	0.345	0.343	0.341	0.339	0.337	0.335
<b>0.6</b>	0.333	0.331	0.329	0.327	0.325	0.323	0.321	0.319	0.317	0.314
<b>0.7</b>	0.312	0.310	0.308	0.306	0.303	0.301	0.299	0.297	0.294	0.292
<b>0.8</b>	0.290	0.287	0.285	0.283	0.280	0.278	0.276	0.273	0.271	0.268
<b>0.9</b>	0.266	0.264	0.261	0.259	0.256	0.254	0.252	0.249	0.247	0.244
<b>1.0</b>	0.242	0.240	0.237	0.235	0.232	0.230	0.227	0.225	0.223	0.220
<b>1.1</b>	0.218	0.215	0.213	0.211	0.208	0.206	0.204	0.201	0.199	0.196
<b>1.2</b>	0.194	0.192	0.190	0.187	0.185	0.183	0.180	0.178	0.176	0.174
<b>1.3</b>	0.171	0.169	0.167	0.165	0.163	0.160	0.158	0.156	0.154	0.152
<b>1.4</b>	0.150	0.148	0.146	0.143	0.141	0.139	0.137	0.135	0.133	0.131
<b>1.5</b>	0.130	0.128	0.126	0.124	0.122	0.120	0.118	0.116	0.114	0.113
<b>1.6</b>	0.111	0.109	0.107	0.106	0.104	0.102	0.101	0.099	0.097	0.096
<b>1.7</b>	0.094	0.092	0.091	0.089	0.088	0.086	0.085	0.083	0.082	0.080
<b>1.8</b>	0.079	0.078	0.076	0.075	0.073	0.072	0.071	0.069	0.068	0.067
<b>1.9</b>	0.066	0.064	0.063	0.062	0.061	0.060	0.058	0.057	0.056	0.055
<b>2.0</b>	0.054	0.053	0.052	0.051	0.050	0.049	0.048	0.047	0.046	0.045
<b>2.1</b>	0.044	0.043	0.042	0.041	0.040	0.040	0.039	0.038	0.037	0.036
<b>2.2</b>	0.035	0.035	0.034	0.033	0.032	0.032	0.031	0.030	0.030	0.029
<b>2.3</b>	0.028	0.028	0.027	0.026	0.026	0.025	0.025	0.024	0.023	0.023
<b>2.4</b>	0.022	0.022	0.021	0.021	0.020	0.020	0.019	0.019	0.018	0.018
<b>2.5</b>	0.018	0.017	0.017	0.016	0.016	0.015	0.015	0.015	0.014	0.014
<b>2.6</b>	0.014	0.013	0.013	0.013	0.012	0.012	0.012	0.011	0.011	0.011
<b>2.7</b>	0.010	0.010	0.009	0.009	0.009	0.009	0.008	0.008	0.008	0.008
<b>2.8</b>	0.007	0.007	0.007	0.007	0.007	0.006	0.006	0.006	0.006	0.006
<b>2.9</b>	0.006	0.005	0.005	0.005	0.005	0.005	0.005	0.004	0.004	0.004
<b>3.0</b>	0.004	0.004	0.004	0.004	0.003	0.003	0.003	0.003	0.003	0.003

Wartości  $t$  są liczbą odchyłeń standardowych od średniej arytmetycznej (wartość 0.00).  
Wartości w tablicy są wyrażone jako frakcja jedności. Od  $t = 2.7$  wartości rozkładu normalnego należy dzielić przez 10.

**Tablica 2**  
**Wartości dystrybuanty rozkładu normalnego**

<b><i>t</i></b>	<b>0.00</b>	<b>0.01</b>	<b>0.02</b>	<b>0.03</b>	<b>0.04</b>	<b>0.05</b>	<b>0.06</b>	<b>0.07</b>	<b>0.08</b>	<b>0.09</b>
<b>0.0</b>	0.500	0.504	0.508	0.512	0.516	0.520	0.524	0.528	0.532	0.536
<b>0.1</b>	0.540	0.544	0.548	0.552	0.556	0.560	0.564	0.568	0.572	0.576
<b>0.2</b>	0.579	0.583	0.587	0.591	0.595	0.599	0.603	0.607	0.611	0.614
<b>0.3</b>	0.618	0.622	0.626	0.630	0.633	0.637	0.641	0.645	0.648	0.652
<b>0.4</b>	0.656	0.659	0.663	0.667	0.670	0.674	0.678	0.681	0.685	0.688
<b>0.5</b>	0.692	0.695	0.699	0.702	0.706	0.709	0.713	0.716	0.719	0.723
<b>0.6</b>	0.726	0.729	0.733	0.736	0.739	0.742	0.746	0.749	0.752	0.755
<b>0.7</b>	0.758	0.761	0.765	0.768	0.771	0.774	0.777	0.780	0.783	0.786
<b>0.8</b>	0.788	0.791	0.794	0.797	0.800	0.803	0.805	0.808	0.811	0.814
<b>0.9</b>	0.816	0.819	0.821	0.824	0.827	0.829	0.832	0.834	0.837	0.839
<b>1.0</b>	0.842	0.844	0.846	0.849	0.851	0.853	0.856	0.858	0.860	0.862
<b>1.1</b>	0.864	0.867	0.869	0.871	0.873	0.875	0.877	0.879	0.881	0.883
<b>1.2</b>	0.885	0.887	0.889	0.891	0.893	0.894	0.896	0.898	0.900	0.902
<b>1.3</b>	0.903	0.905	0.907	0.908	0.910	0.911	0.913	0.915	0.916	0.918
<b>1.4</b>	0.919	0.921	0.922	0.924	0.925	0.926	0.928	0.929	0.930	0.932
<b>1.5</b>	0.933	0.934	0.936	0.937	0.938	0.939	0.940	0.942	0.943	0.944
<b>1.6</b>	0.945	0.946	0.947	0.948	0.949	0.950	0.951	0.952	0.953	0.954
<b>1.7</b>	0.955	0.956	0.957	0.958	0.959	0.960	0.961	0.961	0.962	0.963
<b>1.8</b>	0.964	0.965	0.965	0.966	0.967	0.968	0.968	0.969	0.970	0.970
<b>1.9</b>	0.971	0.972	0.972	0.973	0.973	0.974	0.975	0.975	0.976	0.976
<b>2.0</b>	0.977	0.977	0.978	0.978	0.979	0.979	0.980	0.980	0.981	0.981
<b>2.1</b>	0.982	0.982	0.983	0.983	0.983	0.984	0.984	0.985	0.985	0.985
<b>2.2</b>	0.986	0.986	0.986	0.987	0.987	0.987	0.988	0.988	0.988	0.989
<b>2.3</b>	0.989	0.989	0.989	0.990	0.990	0.990	0.990	0.991	0.991	0.991
<b>2.4</b>	0.991	0.992	0.992	0.992	0.992	0.992	0.993	0.993	0.993	0.993
<b>2.5</b>	0.993	0.993	0.994	0.994	0.994	0.994	0.994	0.994	0.995	0.995
<b>2.6</b>	0.995	0.995	0.995	0.995	0.995	0.995	0.996	0.996	0.996	0.996
<b>2.7</b>	0.996	0.996	0.996	0.996	0.996	0.997	0.997	0.997	0.997	0.997
<b>2.8</b>	0.997	0.997	0.997	0.997	0.997	0.997	0.997	0.997	0.997	0.998
<b>2.9</b>	0.998	0.998	0.998	0.998	0.998	0.998	0.998	0.998	0.998	0.998
<b>3.0</b>	0.998	0.998	0.998	0.998	0.998	0.998	0.998	0.998	0.998	0.998

Wartości  $t$  – patrz tablica 1. Wartości w tablicy podane są jako frakcja pola pod krzywą (całkowitz liczebności) od średniej w górę. Dla wartości poniżej średniej, np. dla  $t = -0.5$ , wartość tablicowa jest równa  $1 - 0.692 = 0.308$ .



**Tablica 3**  
**Wartości funkcji  $t$  Studenta**  
 (prawdopodobieństwa dla testu jedno- lub dwustronnego)

<b>jedno-</b>	<b>0.10</b>	<b>0.05</b>	<b>0.025</b>	<b>0.005</b>	<b>0.0005</b>
<b>dwu-</b>	<b>0.20</b>	<b>0.10</b>	<b>0.05</b>	<b>0.01</b>	<b>0.001</b>
<b>4</b>	1.533	2.132	2.776	4.604	
<b>5</b>	1.476	2.015	2.571	4.032	6.859
<b>6</b>	1.440	1.993	2.447	3.707	5.959
<b>7</b>	1.415	1.895	2.365	3.499	5.405
<b>8</b>	1.397	1.860	2.306	3.355	5.041
<b>9</b>	1.383	1.833	2.262	3.250	4.781
<b>10</b>	1.372	1.812	2.228	3.169	4.587
<b>11</b>	1.363	1.796	2.201	3.106	4.437
<b>12</b>	1.356	1.782	2.179	3.055	4.318
<b>13</b>	1.350	1.771	2.160	3.012	4.221
<b>14</b>	1.345	1.761	2.145	2.977	4.140
<b>15</b>	1.341	1.753	2.131	2.947	4.073
<b>16</b>	1.337	1.746	2.120	2.921	4.015
<b>17</b>	1.333	1.740	2.110	2.898	3.965
<b>18</b>	1.330	1.734	2.101	2.878	3.922
<b>19</b>	1.328	1.729	2.093	2.861	3.883
<b>20</b>	1.325	1.725	2.086	2.845	3.850
<b>21</b>	1.323	1.721	2.080	2.832	3.820
<b>22</b>	1.321	1.717	2.074	2.819	3.792
<b>23</b>	1.319	1.714	2.069	2.808	3.768
<b>24</b>	1.318	1.711	2.064	2.797	3.745
<b>25</b>	1.316	1.708	2.060	2.788	3.725
<b>26</b>	1.315	1.706	2.056	2.779	3.707
<b>27</b>	1.314	1.703	2.052	2.771	3.690
<b>28</b>	1.313	1.701	2.048	2.763	3.674
<b>29</b>	1.311	1.699	2.045	2.756	3.659
<b>30</b>	1.310	1.697	2.042	2.750	3.646
<b>40</b>	1.303	1.684	2.021	2.704	3.551
<b>60</b>	1.296	1.671	2.000	2.660	3.460
<b>90</b>	1.292	1.662	1.987	2.632	3.402
<b>120</b>	1.289	1.658	1.980	2.617	3.373
$\infty$	1.282	1.645	1.960	2.576	3.291

W pierwszej kolumnie podane są liczby stopni swobody (df).

**Tablica 4**  
**Wartości funkcji  $F$  Snedecora**  
 **$P < 0.05$**

<b>df</b>	<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>	<b>6</b>	<b>7</b>	<b>8</b>	<b>9</b>	<b>10</b>
<b>4</b>	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96
<b>5</b>	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74
<b>6</b>	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06
<b>7</b>	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64
<b>8</b>	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35
<b>9</b>	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14
<b>10</b>	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
<b>11</b>	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85
<b>12</b>	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75
<b>13</b>	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67
<b>14</b>	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60
<b>15</b>	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54
<b>16</b>	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49
<b>17</b>	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45
<b>18</b>	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41
<b>19</b>	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38
<b>20</b>	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35
<b>22</b>	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30
<b>24</b>	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25
<b>26</b>	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22
<b>28</b>	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19
<b>30</b>	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16
<b>40</b>	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08
<b>50</b>	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03
<b>60</b>	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99
<b>80</b>	3.96	3.11	2.72	2.49	2.33	2.21	2.13	2.06	2.00	1.95
<b>100</b>	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97	1.93
<b>125</b>	3.90	3.06	2.66	2.43	2.27	2.16	2.07	2.00	1.94	1.89
$\infty$	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83

W pierwszym wierszu podane są liczby stopni swobody (df) dla liczby kontrastów (np. grup), w pierwszej kolumnie – dla błędu (resztowego)

**Tablica 4 (c.d.)**  
**Wartości funkcji  $F$  Snedecora**  
 **$P < 0.01$**

<b>df</b>	<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>	<b>6</b>	<b>7</b>	<b>8</b>	<b>9</b>	<b>10</b>
<b>4</b>	21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7	14.5
<b>5</b>	16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.2	10.1
<b>6</b>	13.7	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87
<b>7</b>	12.2	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62
<b>8</b>	11.3	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81
<b>9</b>	10.6	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26
<b>10</b>	10.0	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85
<b>11</b>	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54
<b>12</b>	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30
<b>13</b>	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10
<b>14</b>	8.86	6.51	5.56	5.04	4.70	4.46	4.28	4.14	4.03	3.94
<b>15</b>	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80
<b>16</b>	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69
<b>17</b>	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59
<b>18</b>	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51
<b>19</b>	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43
<b>20</b>	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37
<b>22</b>	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26
<b>24</b>	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17
<b>26</b>	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09
<b>28</b>	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03
<b>30</b>	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98
<b>40</b>	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80
<b>50</b>	7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89	2.79	2.70
<b>60</b>	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63
<b>80</b>	6.96	4.88	4.04	3.56	3.26	3.04	2.87	2.74	2.64	2.55
<b>100</b>	6.90	4.82	3.98	3.51	3.21	2.99	2.82	2.69	2.59	2.50
<b>125</b>	6.81	4.75	3.92	3.45	3.14	2.92	2.76	2.63	2.53	2.44
$\infty$	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32

**Tablica 5**  
**Wartości funkcji  $\chi^2$  (chi-kwadrat)**

	<b>0.10</b>	<b>0.05</b>	<b>0.01</b>	<b>0.001</b>
<b>1</b>	2.71	3.84	6.63	10.8
<b>2</b>	4.61	5.99	9.21	13.8
<b>3</b>	6.25	7.81	11.4	16.3
<b>4</b>	7.78	9.49	13.3	18.5
<b>5</b>	9.24	11.1	15.1	20.5
<b>6</b>	10.64	12.6	16.8	22.5
<b>7</b>	12.02	14.1	18.5	24.3
<b>8</b>	13.36	15.5	20.1	26.1
<b>9</b>	14.68	16.9	21.7	27.9
<b>10</b>	15.99	18.3	23.2	29.6
<b>11</b>	17.3	19.7	24.7	31.3
<b>12</b>	18.6	21.0	26.2	32.9
<b>13</b>	19.8	22.4	27.7	34.5
<b>14</b>	21.1	23.7	29.1	36.1
<b>15</b>	22.3	25.0	30.6	37.7
<b>16</b>	23.5	26.3	32.0	39.3
<b>17</b>	24.8	27.6	33.4	40.8
<b>18</b>	26.0	28.9	34.8	42.3
<b>19</b>	27.2	30.1	36.2	43.8
<b>20</b>	28.4	31.4	37.6	45.3
<b>25</b>	34.4	37.7	44.3	52.6
<b>30</b>	40.3	43.8	50.9	59.7
<b>40</b>		55.8	63.7	73.4
<b>50</b>		67.5	76.2	86.7

**Tablica 6**  
**Wartości współczynnika korelacji prostej (Pearsona)**

	<b>0.10</b>	<b>0.05</b>	<b>0.01</b>	<b>0.001</b>
<b>4</b>	0.729	0.811	0.917	0.974
<b>5</b>	0.669	0.755	0.875	0.951
<b>6</b>	0.622	0.707	0.834	0.925
<b>7</b>	0.582	0.666	0.798	0.898
<b>8</b>	0.549	0.632	0.765	0.872
<b>9</b>	0.521	0.602	0.735	0.847
<b>10</b>	0.497	0.576	0.708	0.823
<b>11</b>	0.476	0.553	0.636	0.801
<b>12</b>	0.458	0.532	0.661	0.780
<b>13</b>	0.441	0.514	0.641	0.760
<b>14</b>	0.426	0.497	0.623	0.742
<b>15</b>	0.412	0.482	0.606	0.725
<b>16</b>	0.400	0.468	0.590	0.708
<b>17</b>	0.389	0.456	0.575	0.693
<b>18</b>	0.378	0.444	0.561	0.679
<b>19</b>	0.369	0.433	0.549	0.665
<b>20</b>	0.360	0.423	0.537	0.652
<b>22</b>	0.342	0.408	0.517	0.638
<b>24</b>	0.333	0.388	0.500	0.611
<b>26</b>	0.317	0.374	0.480	0.592
<b>28</b>	0.308	0.360	0.463	0.573
<b>30</b>	0.296	0.349	0.449	0.554
<b>35</b>	0.275	0.325	0.418	0.519
<b>40</b>	0.257	0.304	0.393	0.490
<b>45</b>	0.243	0.288	0.372	0.465
<b>50</b>	0.231	0.273	0.354	0.443
<b>60</b>	0.211	0.250	0.325	0.408
<b>70</b>	0.195	0.232	0.302	0.380
<b>80</b>		0.217	0.283	0.357
<b>90</b>		0.205	0.267	0.338
<b>100</b>		0.195	0.254	0.321

**Tablica 7**  
**Wartości współczynnika korelacji wielokrotnej**

<b>k</b>	<b>3</b>		<b>4</b>		<b>5</b>		<b>6</b>	
<b>P</b>	<b>0.05</b>	<b>0.01</b>	<b>0.05</b>	<b>0.01</b>	<b>0.05</b>	<b>0.01</b>	<b>0.05</b>	<b>0.01</b>
<b>6</b>	0.795	0.886	0.839	0.911	0.867	0.927	0.887	0.927
<b>7</b>	0.758	0.855	0.807	0.885	0.838	0.904	0.861	0.904
<b>8</b>	0.726	0.827	0.777	0.860	0.811	0.882	0.836	0.882
<b>9</b>	0.697	0.800	0.750	0.836	0.786	0.861	0.812	0.861
<b>10</b>	0.671	0.776	0.726	0.814	0.763	0.840	0.791	0.840
<b>11</b>	0.648	0.753	0.703	0.793	0.741	0.821	0.769	0.821
<b>12</b>	0.627	0.732	0.683	0.773	0.722	0.802	0.751	0.802
<b>13</b>	0.608	0.712	0.664	0.755	0.703	0.785	0.733	0.785
<b>14</b>	0.590	0.694	0.646	0.737	0.686	0.768	0.716	0.768
<b>15</b>	0.574	0.677	0.630	0.721	0.670	0.752	0.701	0.752
<b>16</b>	0.559	0.662	0.615	0.706	0.655	0.738	0.686	0.738
<b>17</b>	0.545	0.647	0.601	0.691	0.641	0.724	0.672	0.724
<b>18</b>	0.532	0.633	0.587	0.678	0.628	0.710	0.659	0.710
<b>19</b>	0.520	0.620	0.575	0.665	0.615	0.698	0.646	0.698
<b>20</b>	0.509	0.608	0.563	0.652	0.604	0.685	0.635	0.685
<b>22</b>	0.488	0.585	0.542	0.630	0.582	0.663	0.613	0.663
<b>24</b>	0.470	0.565	0.523	0.609	0.562	0.642	0.593	0.642
<b>26</b>	0.454	0.546	0.506	0.590	0.545	0.624	0.576	0.624
<b>28</b>	0.439	0.530	0.490	0.573	0.529	0.606	0.559	0.606
<b>30</b>	0.426	0.514	0.476	0.558	0.514	0.591	0.544	0.591
<b>35</b>	0.397	0.481	0.445	0.523	0.482	0.556	0.511	0.556
<b>40</b>	0.373	0.454	0.419	0.494	0.455	0.526	0.483	0.526
<b>45</b>	0.353	0.430	0.397	0.470	0.432	0.501	0.459	0.501
<b>50</b>	0.336	0.410	0.379	0.449	0.412	0.479	0.439	0.479
<b>60</b>	0.308	0.377	0.348	0.414	0.380	0.442	0.405	0.442
<b>70</b>	0.286	0.351	0.324	0.386	0.354	0.413	0.378	0.413
<b>80</b>	0.269	0.330	0.304	0.362	0.332	0.389	0.354	0.389
<b>90</b>	0.254	0.312	0.288	0.343	0.315	0.368	0.337	0.368
<b>100</b>	0.241	0.297	0.274	0.327	0.300	0.351	0.321	0.351

**k** – liczba zmiennych; **P** – poziom prawdopodobieństwa

**Tablica 8**  
**Wartości funkcji  $G = n \cdot \ln(n)$**

	0	1	2	3	4	5	6	7	8	9
<b>0</b>	0	0	1.39	3.30	5.55	8.05	10.75	13.62	16.64	19.78
<b>10</b>	23.03	26.38	29.82	33.34	36.95	40.62	44.36	48.16	52.03	55.94
<b>20</b>	59.91	63.93	68.00	72.12	76.27	80.47	84.71	88.99	93.30	97.65
<b>30</b>	102.04	106.45	110.90	115.38	119.90	124.44	129.01	133.60	138.23	142.88
<b>40</b>	147.56	152.26	156.98	161.73	166.50	171.30	176.12	180.96	185.82	190.70
<b>50</b>	195.60	200.52	205.46	210.43	215.41	220.40	225.42	230.45	235.51	240.57
<b>60</b>	245.66	250.76	255.88	261.02	266.17	271.34	276.52	281.71	286.93	292.15
<b>70</b>	297.39	302.65	307.92	313.20	318.50	323.81	329.14	334.47	339.82	345.19
<b>80</b>	350.56	355.95	361.35	366.76	372.19	377.63	383.07	388.53	394.01	399.49
<b>90</b>	404.98	410.49	416.00	421.53	427.07	432.62	438.18	443.75	449.33	454.92
<b>100</b>	460.52	466.13	471.75	477.38	483.02	488.67	494.32	499.99	505.67	511.36
<b>110</b>	517.05	522.76	528.47	534.19	539.93	545.67	551.42	557.17	562.94	568.72
<b>120</b>	574.50	580.29	586.09	591.90	597.71	603.54	609.37	615.21	621.06	626.92
<b>130</b>	632.78	638.65	644.53	650.42	656.31	662.21	668.12	674.04	679.96	685.89
<b>140</b>	691.83	697.78	703.73	709.69	715.65	721.63	727.61	733.59	739.59	745.59
<b>150</b>	751.60	757.61	763.63	769.66	775.69	781.73	787.78	793.83	799.89	805.96
<b>160</b>	812.03	818.11	824.19	830.28	836.38	842.48	848.59	854.70	860.83	866.95
<b>170</b>	873.09	879.22	885.37	891.52	897.68	903.84	910.01	916.18	922.36	928.54
<b>180</b>	934.73	940.93	947.13	953.34	959.55	965.77	971.99	978.22	984.45	990.69
<b>190</b>	996.93	1003.18	1009.44	1015.70	1021.96	1028.23	1034.51	1040.79	1047.08	1053.37
<b>200</b>	1059.66	1065.96	1072.27	1078.58	1084.90	1091.22	1097.54	1103.87	1110.21	1116.55
<b>210</b>	1122.89	1129.24	1135.60	1141.96	1148.32	1154.69	1161.06	1167.44	1173.82	1180.21
<b>220</b>	1186.60	1192.99	1199.39	1205.80	1212.21	1218.62	1225.04	1231.46	1237.89	1244.32
<b>230</b>	1250.76	1257.20	1263.64	1270.09	1276.55	1283.00	1289.46	1295.93	1302.40	1308.87
<b>240</b>	1315.35	1321.84	1328.32	1334.81	1341.31	1347.81	1354.31	1360.82	1367.33	1373.85
<b>250</b>	1380.37	1386.89	1393.42	1399.95	1406.48	1413.02	1419.57	1426.11	1432.66	1439.22
<b>260</b>	1445.78	1452.34	1458.91	1465.48	1472.05	1478.63	1485.21	1491.80	1498.38	1504.98
<b>270</b>	1511.57	1518.17	1524.78	1531.39	1538.00	1544.61	1551.23	1557.85	1564.48	1571.11
<b>280</b>	1577.74	1584.38	1591.02	1597.66	1604.31	1610.96	1617.61	1624.27	1630.93	1637.60
<b>290</b>	1644.27	1650.94	1657.61	1664.29	1670.97	1677.66	1684.35	1691.04	1697.73	1704.43
<b>300</b>	1711.13	1717.84	1724.55	1731.26	1737.98	1744.70	1751.42	1758.14	1764.87	1771.60
<b>310</b>	1778.34	1785.08	1791.82	1798.56	1805.31	1812.06	1818.81	1825.57	1832.33	1839.10
<b>320</b>	1845.86	1852.63	1859.41	1866.18	1872.96	1879.74	1886.53	1893.32	1900.11	1906.90
<b>330</b>	1913.70	1920.50	1927.30	1934.11	1940.92	1947.73	1954.55	1961.37	1968.19	1975.01
<b>340</b>	1981.84	1988.67	1995.51	2002.34	2009.18	2016.02	2022.87	2029.72	2036.57	2043.42
<b>350</b>	2050.28	2057.14	2064.00	2070.86	2077.73	2084.60	2091.48	2098.35	2105.23	2112.11
<b>360</b>	2119.00	2125.88	2132.78	2139.67	2146.56	2153.46	2160.36	2167.27	2174.17	2181.08
<b>370</b>	2188.00	2194.91	2201.83	2208.75	2215.67	2222.60	2229.53	2236.46	2243.39	2250.33
<b>380</b>	2257.27	2264.21	2271.15	2278.10	2285.05	2292.00	2298.95	2305.91	2312.87	2319.83
<b>390</b>	2326.80	2333.76	2340.73	2347.71	2354.68	2361.66	2368.64	2375.62	2382.61	2389.60
<b>400</b>	2396.59	2403.58	2410.57	2417.57	2424.57	2431.57	2438.58	2445.59	2452.60	2459.61
<b>410</b>	2466.62	2473.64	2480.66	2487.68	2494.71	2501.74	2508.77	2515.80	2522.83	2529.87
<b>420</b>	2536.91	2543.95	2550.99	2558.04	2565.09	2572.14	2579.19	2586.25	2593.30	2600.37
<b>430</b>	2607.43	2614.49	2621.56	2628.63	2635.70	2642.78	2649.85	2656.93	2664.01	2671.10

**Tablica 9**  
**Wartości funkcji**  $z = \frac{1}{2} \ln \left( \frac{1+r}{1-r} \right)$

<b>r</b>	<b>0.00</b>	<b>0.01</b>	<b>0.02</b>	<b>0.03</b>	<b>0.04</b>	<b>0.05</b>	<b>0.06</b>	<b>0.07</b>	<b>0.08</b>	<b>0.09</b>
<b>0.0</b>	0.000	0.005	0.010	0.015	0.020	0.026	0.031	0.036	0.042	0.047
<b>0.1</b>	0.053	0.059	0.064	0.070	0.076	0.082	0.088	0.095	0.101	0.107
<b>0.2</b>	0.114	0.121	0.127	0.134	0.142					
	<b>0.000</b>	<b>0.003</b>	<b>0.005</b>	<b>0.008</b>	<b>0.010</b>	<b>0.013</b>	<b>0.015</b>	<b>0.018</b>	<b>0.020</b>	<b>0.023</b>
<b>0.25</b>	0.149	0.151	0.152	0.155	0.156	0.158	0.160	0.162	0.164	0.166
<b>0.30</b>	0.187	0.190	0.192	0.194	0.196	0.198	0.200	0.202	0.204	0.207
<b>0.35</b>	0.231	0.234	0.236	0.238	0.240	0.243	0.245	0.248	0.250	0.253
<b>0.40</b>	0.280	0.284	0.286	0.289	0.291	0.294	0.297	0.300	0.302	0.306
<b>0.45</b>	0.338	0.342	0.344	0.348	0.350	0.354	0.357	0.361	0.363	0.367
<b>0.50</b>	0.405	0.410	0.413	0.417	0.421	0.425	0.428	0.433	0.436	0.441
<b>0.55</b>	0.487	0.492	0.496	0.502	0.505	0.511	0.515	0.521	0.525	0.530
<b>0.60</b>	0.588	0.594	0.599	0.606	0.611	0.618	0.623	0.630	0.635	0.642
<b>0.65</b>	0.715	0.724	0.730	0.739	0.745	0.755	0.761	0.771	0.777	0.787
<b>0.70</b>	0.884	0.896	0.904	0.917	0.925	0.938	0.946	0.960		
	<b>0.000</b>	<b>0.002</b>	<b>0.004</b>	<b>0.006</b>	<b>0.008</b>	<b>0.010</b>	<b>0.012</b>	<b>0.014</b>	<b>0.016</b>	<b>0.018</b>
<b>0.72</b>	0.968	0.977	0.987	0.996	1.005	1.015	1.025	1.035	1.045	1.055
<b>0.74</b>	1.065	1.076	1.086	1.097	1.108	1.119	1.131	1.142	1.154	1.166
<b>0.76</b>	1.178	1.190	1.203	1.215	1.228	1.241	1.255	1.268	1.282	1.296
<b>0.78</b>	1.310	1.325	1.340	1.355	1.371	1.386	1.402	1.419	1.435	1.452
<b>0.80</b>	1.469	1.487	1.505	1.523	1.542	1.561	1.581	1.601	1.621	1.642
<b>0.82</b>	1.663	1.685	1.707	1.730	1.754	1.777	1.802	1.827	1.852	1.879
<b>0.84</b>	1.906	1.933	1.961	1.990	2.020	2.051	2.082	2.114	2.147	2.181
<b>0.86</b>	2.216	2.252	2.289	2.328	2.367	2.407	2.449	2.492	2.537	2.583
<b>0.88</b>	2.630	2.679	2.730	2.783	2.837	2.894	2.952	3.013	3.076	3.141
	<b>0.000</b>	<b>0.001</b>	<b>0.002</b>	<b>0.003</b>	<b>0.004</b>	<b>0.005</b>	<b>0.006</b>	<b>0.007</b>	<b>0.008</b>	<b>0.009</b>
<b>0.90</b>	3.209	3.244	3.280	3.317	3.354	3.392	3.431	3.471	3.511	3.553
<b>0.91</b>	3.595	3.638	3.683	3.728	3.774	3.822	3.870	3.920	3.971	4.023
<b>0.92</b>	4.077	4.132	4.188	4.246	4.305	4.366	4.429	4.493	4.559	4.627
<b>0.93</b>	4.697	4.768	4.842	4.918	4.997	5.078	5.161	5.247	5.336	5.428
<b>0.94</b>	5.522	5.620	5.722	5.827	5.935	6.048	6.165	6.286	6.412	6.542
<b>0.95</b>	6.678	6.820	6.967	7.121	7.281	7.449	7.624	7.807	7.999	8.200
<b>0.96</b>	8.412	8.634	8.868	9.115	9.375	9.650	9.941	10.25	10.57	10.92
<b>0.97</b>	11.30	11.69	12.12	12.58	13.07	13.61	14.18	14.81	15.50	16.25
<b>0.98</b>	17.07	17.99	19.00	20.13	21.41	22.85	24.50	26.40	28.63	31.25
<b>0.99</b>	34.40	38.25	43.07	49.26	57.51	69.06	86.39	115.2	173.0	346.3





## SKOROWIDZ

- analiza regresji liniowej 59  
 wariacji, jednokierunkowa 35  
 wariacji, dwukierunkowa 41  
 wariacji, wielocechowa 74  
 wielocechowa 71
- ankieta 13
- ANOVA, patrz analiza wariacji
- błąd I rodzaju 50, 91  
 II rodzaju 50, 91  
 standardowy 24  
 standardowy różnicy 33  
 w regresji 57 – 59
- brakujące dane, wstawianie 47
- cecha - patrz zmienna 9
- centyle 94
- cyfry znaczące 9
- dane niezależne 33, 47, 100  
 zależne 34, 47, 100
- dystrybuanta 18 – 20, 90 – 94
- formuły, tworzenie w EXCELu 101
- funkcja chi-kwadrat 83, 95  
 cykliczna 63  
*F* 36  
*G* 84  
 minimaksowa 63  
 monotoniczna 63
- funkcje statystyczne 99
- hipoteza zerowa 49
- histogram 14 – 16
- interakcja 42 – 44
- jednorodność statystyczna 16, 90
- jednorodność stochastyczna 16, 90
- korelacja 52  
 cząstkowa 56, 80  
 dwu zmiennych 52  
 krzywoliniowa, patrz stosunek korelacyjny  
 Spearmana 87  
 wielokrotna 76
- korelogram 53
- krzywa Gaussa 17
- kwestionariusz 13
- liczebności skumulowane 18
- mediana 23, 25
- miara przeciętna, patrz miara centralna 23
- modalna 23, 25
- najmniejsza istotna różnica 37
- norma 89
- normalizacja danych 26
- normowanie, patrz zmienna unormowana
- odchylenia normalne 20
- odchylenie standardowe 17, 23
- odpowiedzi stopniowane 14  
 wielokrotnego wyboru 13  
 zero-jedynkowe 13
- odrzućanie danych 49, 54
- ortogonalność 41
- poprawka Shepparda 26
- próba losowa 11  
 reprezentatywna 11  
 warstwowa 11
- profil wielocechowy 71 – 74
- przedziały klasowe 14, 15, 18  
 referencyjne 90, 91
- rangowanie 87
- ranking 71
- regresja 52  
 krzywoliniowa 63  
 liniowa 56  
 paraboliczna 66
- równanie regresji 56
- rozkład danych, badanie 19, 99  
 dwuszczytowy 16  
 jednoszczytowy 15  
 niejednorodny 20  
 normalny 15, 17, 20  
 prostokątny 16  
 skośny 15, 16, 20
- rozstęp 24
- SD, patrz odchylenie standardowe
- SE, patrz błąd standardowy
- skala naturalna 9  
 otwarta 9  
 sztuczna 9  
 zamknięta 9, 29
- średnia arytmetyczna 17, 23, 25, 99  
 ruchoma 68  
 ważona 31
- stała regresji 56
- stopnie swobody 24

stosunek korelacyjny 68, 84  
szereg rozdzielczy 14  
test     Bartletta 37  
          dwustronny 50, 100  
          jednostronny 50, 100  
          Scheffé'go 37  
           $t$  (Studenta) 33, 34, 100  
          topologiczny 85  
testy    nieparametryczne 83  
          *post-hoc* 36  
          statystyczne 19, 49  
transformacja kątowna 30  
          logarytmiczna 26, 64  
          logitowa 30  
          pierwiastkowa 28  
          danych 26  
układ    bloków losowych 44  
          *cross-over* 39  
          hierarchiczno-krzyżowy 45  
          krzyżowy 41  
waga, w układach nieortogonalnych 43  
wagi danych unormowanych 73  
wariancja 24  
wartości referencyjne 89  
współczynnik korelacji 52, 100  
          regresji 56  
          zmienności 15, 26  
współczynniki korelacji, porównywanie 55  
zmienna ciągła 9  
          dyskretna 9  
          unormowana 30, 71 – 74  
zredukowana suma kwadratów 40

**Literatura uzupełniająca**

C.I.Bliss

Statistics in biology, vol. I - III

McGraw-Hill, N.Y., 1967

G.F.Ferguson, Y.Takane

Analiza statystyczna w psychologii i pedagogice

PWN, 1999

D.F.Morrison

Wielowymiarowa analiza statystyczna

PWN, 1990

R.R.Sokal, F.J.Rohlf

Biometry (3<sup>rd</sup> ed.)

W.H.Freeman, N.Y., 1998

R.Stupnicki

Analiza i prezentacja danych ankietowych

Wyd. AWF Warszawa, 2015

<https://platforma.awf.edu.pl/Materials/Details/NTU1YjFkODZIMWY0NjEyY2QyZDU5Nz>

[Rh](#)