

Zaufaj posiadanym danym!

Conrad Carlberg

Analiza statystyczna. Microsoft® Excel 2010 PL

$x A 3 * 7 B$

3452717782712617787
3452717782712617787

$b * c 12$

000717782712617787
34527175612617787
34527175612617787

45%

2345
3452900000

33%

345271712617787

Tytuł oryginału: Statistical Analysis: Microsoft Excel 2010

Tłumaczenie: Maria Chaniewska

ISBN: 978-83-246-3668-6

Authorized translation from the English language edition, entitled: Statistical Analysis: Microsoft Excel 2010, First Edition; ISBN 9780789747204, by Conrad Carlberg, published by Pearson Education, Inc, publishing as QUE Publishing, Copyright © 2011 by Pearson Education, Inc.

All rights reserved. No part of this book may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording or by any information storage retrieval system, without permission from Pearson Education Inc.

Polish language edition published by Helion S.A.
Copyright © 2012.

Wszelkie prawa zastrzeżone. Nieautoryzowane rozpowszechnianie całości lub fragmentu niniejszej publikacji w jakiejkolwiek postaci jest zabronione. Wykonywanie kopii metodą kserograficzną, fotograficzną, a także kopiowanie książki na nośniku filmowym, magnetycznym lub innym powoduje naruszenie praw autorskich niniejszej publikacji.

Wszystkie znaki występujące w tekście są zastrzeżonymi znakami firmowymi bądź towarowymi ich właścicieli.

Autor oraz Wydawnictwo HELION dołożyli wszelkich starań, by zawarte w tej książce informacje były kompletne i rzetelne. Nie biorą jednak żadnej odpowiedzialności ani za ich wykorzystanie, ani za związane z tym ewentualne naruszenie praw patentowych lub autorskich. Autor oraz Wydawnictwo HELION nie ponoszą również żadnej odpowiedzialności za ewentualne szkody wynikłe z wykorzystania informacji zawartych w książce.

Wydawnictwo HELION
ul. Kościuszki 1c, 44-100 GLIWICE
tel. 32 231 22 19, 32 230 98 63
e-mail: helion@helion.pl
WWW: <http://helion.pl> (księgarnia internetowa, katalog książek)

Drogi Czytelniku!

Jeżeli chcesz ocenić tę książkę, zajrzyj pod adres

<http://helion.pl/user/opinie/anaste>

Możesz tam wpisać swoje uwagi, spostrzeżenia, recenzję.

Printed in Poland.

- [Kup książkę](#)
- [Poleć książkę](#)
- [Oceń książkę](#)

- [Księgarnia internetowa](#)
- [Lubię to! » Nasza społeczność](#)

Spis treści

Wstęp	11
Stosowanie Excela do analizy statystycznej	11
Czytelniczy i Excel	12
Porządkowanie terminów	13
Upraszczanie spraw	14
Zły produkt?	15
Odwracanie kota ogonem	17
Co zawiera książka?	17
1 Zmienne i wartości	19
Zmienne i wartości	19
Zapisywanie danych w postaci list	20
Skale pomiarowe	23
Skale nominalne	23
Skale liczbowe	25
Określanie wartości przedziałowych na podstawie wartości tekstowych	26
Graficzna prezentacja zmiennych liczbowych w Excelu	29
Graficzna prezentacja dwóch zmiennych	29
Pojęcie rozkładów liczebności	31
Stosowanie rozkładów liczebności	34
Budowanie rozkładu liczebności na podstawie próby	37
Tworzenie symulowanych rozkładów liczebności	45
2 Jak się skupiają wartości	49
Obliczanie średniej arytmetycznej	51
Funkcje, argumenty i wyniki	51
Formuły, wyniki i formaty	54
Minimalizowanie rozproszenia	56
Obliczanie mediany	61
Decyzja o użyciu mediany	63
Obliczanie wartości modalnej	63
Otrzymywanie wartości modalnej kategorii za pomocą formuły	69
Od tendencji centralnej do rozrzutu	75
3 Rozrzut — jak się rozpraszają wartości	77
Mierzenie rozproszenia za pomocą rozstępu	78
Koncepcja odchylenia standardowego	81
Dopasowanie do standardu	81
Myślenie w kategoriach odchyłeń standardowych	83

Obliczanie odchylenia standardowego i wariancji	85
Podnoszenie odchyleń do kwadratu	88
Parametry populacji i przykładowe statystyki	88
Dzielenie przez N-1	89
Obciążenie estymacji	91
Liczba stopni swobody	92
Funkcje Excela do mierzenia rozproszenia	93
Funkcje odchylenia standardowego	93
Funkcje wariancji	94
4 Jak zmienne wspólnie się zmieniają — korelacja	95
Pojęcie korelacji	95
Wyznaczanie współczynnika korelacji	97
Korzystanie z funkcji WSP.KORELACJI()	103
Korzystanie z narzędzi analitycznych	107
Korzystanie z narzędzia Korelacja	108
Korelacja nie oznacza przyczynowości	111
Stosowanie korelacji	113
Usuwanie efektów skali	114
Korzystanie z funkcji Excela	117
Prognozowanie wartości	118
Szacowanie funkcji regresji	120
Stosowanie funkcji REGLINW() do regresji wielorakiej	123
Łączenie predyktorów	123
Najlepsza kombinacja liniowa	124
Pojęcie współdzielonej zmienności	127
Uwaga techniczna: algebra macierzowa i regresja wieloraka w Excelu	129
Przechodzenie do wnioskowania statystycznego	131
5 Jak zmienne są wspólnie klasyfikowane — tabele kontyngencji	133
Jednowymiarowe tabele przestawne	133
Przeprowadzanie testu statystycznego	136
Stawianie założeń	141
Wybór losowy	141
Niezależność wyborów	143
Wzór na prawdopodobieństwo w rozkładzie dwumianowym	144
Korzystanie z funkcji ROZKŁ.DWUM.ODWR()	145
Dwuwymiarowe tabele przestawne	151
Prawdopodobieństwa i zdarzenia niezależne	154
Sprawdzanie niezależności klasyfikacji	156
Efekt Yule'a i Simpsona	162
Podsumowanie funkcji χ^2	164

6	Prawdopodobieństwo statystyki	173
	Problemy z dokumentacją Excela	173
	Kontekst wnioskowania statystycznego	175
	Pojęcie trafności wewnętrznej	176
	Test F z dwiema próbami dla wariancji	181
	Po co przeprowadzać ten test?	182
7	Praca z rozkładem normalnym w Excelu	193
	Opis rozkładu normalnego	193
	Charakterystyki rozkładu normalnego	193
	Standaryzowany rozkład normalny	199
	Funkcje Excela dla rozkładu normalnego	200
	Funkcja ROZKŁ.NORMALNY()	200
	Funkcja ROZKŁ.NORMALNY.ODWR()	203
	Przedziały ufności i rozkład normalny	205
	Znaczenie przedziału ufności	206
	Konstruowanie przedziału ufności	207
	Funkcje arkusza Excela, które wyznaczają przedziały ufności	210
	Korzystanie z funkcji UFNOŚĆ.NORM() i UFNOŚĆ()	211
	Korzystanie z funkcji UFNOŚĆ.T()	214
	Zastosowanie dodatku Analiza danych do przedziałów ufności	215
	Przedziały ufności i testowanie hipotez	217
	Centralne twierdzenie graniczne	217
	Upraszczanie spraw	219
	Ulepszanie spraw	221
8	Testowanie różnic pomiędzy średnimi — podstawy	223
	Testowanie średnich — przesłanki	224
	Stosowanie testu z	225
	Stosowanie błędu standardowego średniej	228
	Tworzenie wykresów	232
	Stosowanie testu t zamiast testu z	240
	Definiowanie reguły decyzyjnej	242
	Pojęcie mocy statystycznej	246
9	Testowanie różnic pomiędzy średnimi — dalsze zagadnienia	253
	Stosowanie funkcji Excela ROZKŁ.T() i ROZKŁ.T.ODWR() do weryfikacji hipotez	254
	Stawianie hipotez jednostronnych i dwustronnych	254
	Wybieranie funkcji rozkładu t-Studenta w Excelu na podstawie hipotez	255
	Uzupełnienie obrazu za pomocą funkcji ROZKŁ.T()	264
	Korzystanie z funkcji T.TEST()	265
	Stopnie swobody w funkcjach Excela	265
	Równe i nierówne liczebności grup	266
	Składnia funkcji T.TEST()	269

Korzystanie z narzędzi do testów t dodatku Analiza danych	282
Wariancje grup w testach t	283
Wizualizacja mocy statystycznej	288
Kiedy unikać testów t	289
10 Testowanie różnic pomiędzy średnimi — analiza wariancji	291
Dlaczego nie należy stosować testów t?	292
Koncepcja analizy wariancji	293
Dzielenie wyników	294
Porównywanie wariancji	297
Test F	301
Stosowanie funkcji F arkusza Excela	305
Korzystanie z funkcji ROZKŁ.F() i ROZKŁ.F.PS()	305
Korzystanie z funkcji ROZKŁ.F.ODWR() i ROZKŁAD.F.ODW()	306
Rozkład F	308
Nierówne liczebności grup	309
Procedury porównań wielokrotnych	311
Procedura Scheffého	312
Planowane różnice ortogonalne	317
11 Analiza wariancji — dalsze zagadnienia	321
Czynnikowa analiza wariancji	321
Inne przesłanki dla zastosowania wielu czynników	323
Korzystanie z narzędzia do dwuczynnikowej analizy wariancji	325
Znaczenie interakcji	328
Istotność statystyczna interakcji	329
Obliczanie efektu interakcji	330
Problem nierównych liczebności grup	335
Powtarzane obserwacje — analiza dwuczynnikowa bez powtórzeń	338
Funkcje i narzędzia Excela — ograniczenia i rozwiązania	339
Moc testu F	341
Modele mieszane	342
12 Analiza regresji wielorakiej i rekodowanie zmiennych nominalnych — podstawy	343
Regresja wieloraka a analiza wariancji	344
Stosowanie rekodowania zmiennych	346
Rekodowanie zmiennych — ogólne zasady	347
Inne typy kodowania	348
Regresja wieloraka i proporcje wariancji	349
Gładkie przejście od analizy wariancji do regresji	352
Znaczenie rekodowania zmiennych	355

Rekodowanie zmiennych w Excelu	357
Korzystanie z narzędzia Regresja w Excelu do analizy grup o nierównych liczebnościach	360
Rekodowanie zmiennych, regresja i schematy czynnikowe w Excelu	362
Stosowanie kontroli statystycznej z korelacjami semicząstkowymi	364
Stosowanie kwadratów współczynników korelacji semicząstkowej do otrzymania prawidłowej sumy kwadratów	366
Stosowanie funkcji REGLINW() zamiast kwadratów współczynników korelacji semicząstkowej	367
Praca z resztami	369
Stosowanie bezwzględnego i względnego adresowania Excela do wyznaczania kwadratów współczynników korelacji semicząstkowej	372
13 Analiza regresji wielorakiej — dalsze zagadnienia	377
Analiza nierównoważonych schematów czynnikowych za pomocą regresji wielorakiej	378
W schemacie zrównoważonym zmienne nie są skorelowane	379
W schemacie nierównoważonym zmienne są skorelowane	380
Kolejność wpisów w schemacie zrównoważonym nie jest istotna	381
Kolejność wpisów w schemacie nierównoważonym jest istotna	384
Wahające się części wariancji	386
Schematy eksperymentalne, badania obserwacyjne i korelacja	387
Korzystanie z wszystkich statystyk funkcji REGLINP()	390
Wykorzystanie współczynników regresji	391
Wykorzystanie błędów standardowych	392
Wykorzystanie wyrażu wolnego	392
Trzeci, czwarty i piąty wiersz wyników funkcji REGLINP()	393
Radzenie sobie z nierównymi liczebnościami grup w prawdziwym eksperymencie	397
Radzenie sobie z nierównymi liczebnościami grup w badaniach obserwacyjnych	399
14 Analiza kowariancji — podstawy	403
Cele analizy kowariancji	404
Większa moc	404
Redukcja obciążenia	405
Stosowanie analizy kowariancji w celu zwiększenia mocy statystycznej	405
Analiza wariancji nie znajduje znaczącej różnicy średnich	406
Dodawanie zmiennej towarzyszącej do analizy	408
Testowanie średniego współczynnika regresji	416
Usuwanie obciążenia — inny wynik	418
15 Analiza kowariancji — dalsze zagadnienia	425
Korygowanie średnich za pomocą funkcji REGLINP() i rekodowania zmiennych	425
Rekodowanie zmiennych za pomocą zmiennych wskaźnikowych i skorygowane średnie grup	431

Wielokrotne porównania po analizie kowariancji	434
Stosowanie metody Scheffégo	434
Stosowanie planowanych różnic	439
Analiza kowariancji wielorakiej	441
Decyzja o zastosowaniu wielu zmiennych towarzyszących	442
Dwie zmienne towarzyszące — przykład	443
Skorowidz	447

Praca z rozkładem normalnym w Excelu

7

Opis rozkładu normalnego

Nie da się przejść przez życie bez prawie codziennego kontaktu z rozkładem normalnym, czyli krzywą dzwonową. Twoje oceny w szkole podstawowej i średniej były umieszczane „na krzywej”. Wzrost i waga osób w rodzinie, sąsiedztwie i kraju są zbliżone do krzywej normalnej. Liczba wypadnięć orła podczas 10 rzutów symetryczną monetą przypomina krzywą normalną. Nawet ta mocno skrócona lista ilustruje niezwykłość fenomenu, który zaczęto dostrzegać 300 lat temu.

Rozkład normalny zajmuje specjalne miejsce w teorii statystyki i rachunku prawdopodobieństwa. To główny powód, dla którego Excel oferuje więcej funkcji arkusza dotyczących rozkładu normalnego niż dowolnego innego, na przykład t-Studenta, dwumianowego, Poissona itd. Innym powodem przykładania tak dużej wagi do rozkładu normalnego jest to, że tak wiele zmiennych, które interesują badaczy — oprócz paru już wspomnianych — ma rozkład normalny.

Charakterystyki rozkładu normalnego

Nie istnieje tylko jeden rozkład normalny, ale ich nieskończona liczba. Chociaż jest ich tak dużo, nigdy nie spotkasz żadnego z nich w naturze.

W TYM ROZDZIALE:

Opis rozkładu normalnego	193
Funkcje Excela dla rozkładu normalnego ..	200
Przedziały ufności i rozkład normalny	205
Centralne twierdzenie graniczne	217



Nie są to sprzeczne stwierdzenia. Istnieje krzywa normalna — lub, jak wolisz, rozkład normalny, lub krzywa dzwonowa, lub krzywa Gaussa — dla każdej pary liczb, ponieważ krzywa normalna może mieć dowolną średnią i dowolne odchylenie standardowe. Krzywa normalna może mieć średnią równą 100 i odchylenie standardowe równe 16 lub średnią 54,3 i odchylenie standardowe równe 10. Wszystko zależy od mierzonej zmiennej.

Nigdy jednak nie zobaczysz rozkładu normalnego w naturze, ponieważ natura jest nieuporządkowana. Widzisz dużą liczbę zmiennych, których rozkłady bardzo przypominają rozkład normalny. Jednak rozkład normalny jest wynikiem równania i dlatego może być dokładnie wykreślony. Jeżeli spróbujesz emulować krzywą normalną, przedstawiając wykres osób, których wzrost to 142, 143 cm itd., dostrzeżesz rozkład przypominający krzywą normalną już wtedy, gdy przedstawisz na wykresie dane około 30 osób.

Gdy liczebność próby losowej osiągnie setki, zobaczysz, że rozkład liczebności wygląda w przybliżeniu normalnie — nie całkowicie, ale dość blisko. Gdy będą to tysiące, zobaczysz, że rozkład liczebności nie jest wizualnie rozróżnialny od krzywej normalnej. Ale jeżeli zastosujesz funkcje dla skośności i kurtozy opisane w tym rozdziale, zauważysz, że ta krzywa po prostu nie jest doskonale normalna. Po pierwsze, wpływają na to drobne błędy próbkowania, po drugie, pomiary nie są doskonale dokładne, a po trzecie i najważniejsze, rozkład normalny obejmuje wszystkie wartości rzeczywiste, a w przypadku każdego obserwowanego zjawiska wartości są ograniczone z góry i z dołu.

Skośność

Rozkład normalny nie jest skośny lewo- ani prawostronnie, ale jest symetryczny. Skośne rozkłady mają wartości, których częstości skupiają się z jednej strony i rozciągają z drugiej.

Skośność a odchylenia standardowe

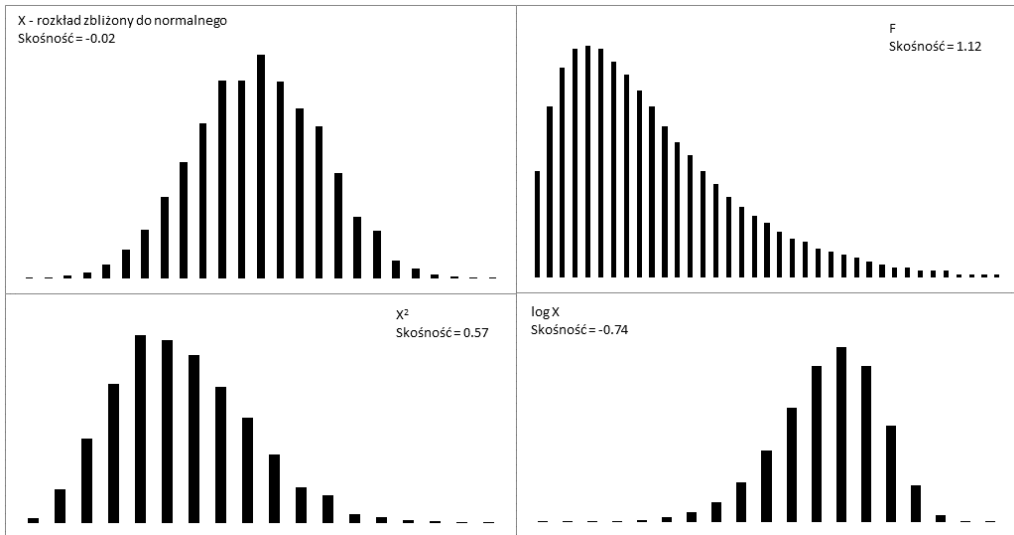
Asymetria skośnego rozkładu powoduje, że znaczenie odchylenia standardowego jest inne od jego znaczenia w rozkładzie symetrycznym, takim jak krzywa normalna lub rozkład t-Studenta (informacje na temat rozkładu t-Studenta znajdziesz w rozdziałach 8. i 9.). W rozkładzie symetrycznym, na przykład normalnym, blisko 34% pola powierzchni obszaru pod krzywą znajduje się pomiędzy średnią a jednym odchyleniem standardowym poniżej średniej. Ponieważ ten rozkład jest symetryczny, dodatkowe 34% pola powierzchni znajduje się również pomiędzy średnią a jednym odchyleniem standardowym powyżej średniej.

Jednak asymetria rozkładu skośnego powoduje, że równe procenty w rozkładzie symetrycznym stają się nierówne. Na przykład w rozkładzie, który jest skośny prawostronnie, możesz znaleźć 45% pola powierzchni pod krzywą pomiędzy średnią arytmetyczną a jednym standardowym odchyleniem poniżej tej średniej. Inne 25% może się znajdować pomiędzy średnią a jednym standardowym odchyleniem powyżej niej.

W tym przypadku nadal masz około 68% pola powierzchni pomiędzy jednym standardowym odchyleniem poniżej i jednym standardowym odchyleniem powyżej średniej arytmetycznej. Jednak to 68% jest podzielone tak, że jego duża część znajduje się poniżej średniej.

Graficzne przedstawianie rozkładów skośnych

Na rysunku 7.1 przedstawiono kilka rozkładów z różnymi stopniami skośności.



Rysunek 7.1. Krzywa jest określona jako skośna w kierunku, w którym się rozciąga — krzywa $\log X$ jest skośna lewostronnie albo skośna ujemnie

Krzywa normalna pokazana na rysunku 7.1 (oparta na wygenerowanej za pomocą dodatku *Analiza danych* Excela losowej próbie 5000 liczb) nie jest idealną krzywą normalną, ale jej bliską aproksymacją. Jej skośność obliczona przez funkcję Excela `SKOŚNOŚĆ()` wynosi $-0,02$. Jest to wartość bardzo bliska zera. Doskonała krzywa normalna ma skośność równą dokładnie 0.

Krzywe X^2 i $\log X$ na rysunku 7.1 są oparte na tych samych wartościach X co ilustracja przybliżonego rozkładu normalnego. Krzywa X^2 rozciąga się w prawo i ma dodatnią skośność równą 0,57. Krzywa $\log X$ rozciąga się w lewo i ma skośność ujemną równą $-0,74$. Ogólnie rzecz biorąc, ujemne miary skośności wskazują na rozkład, który rozciąga się w lewo, a dodatnie miary skośności dotyczą wykresów rozciągniętych w prawo.

Krzywa F na rysunku 7.1 jest oparta na prawdziwym rozkładzie F z 4 i 100 stopniami swobody. (Więcej na temat rozkładów F znajdziesz w tej książce, począwszy od rozdziału 10., „Testowanie różnic pomiędzy średnimi — analiza wariancji”. Rozkład F jest oparty na porcji dwóch wariancji, z których każda ma pewną liczbę stopni swobody). Rozkłady F zawsze są skośne prawostronnie. Ten został tu umieszczony, abyś mógł porównać go z innym ważnym rozkładem — t-Studenta, który pojawi się w następnym podpunkcie na temat kurtozy krzywej.

Obliczanie skośności

Istnieje kilka metod obliczania skośności zbioru liczb. Chociaż zwracane wartości są bliskie sobie nawzajem, żadne dwie metody nie dają dokładnie tych samych wyników. Niestety naukowcy nie doszli w tej sprawie do porozumienia. Piszę tu o większości stosowanych wzorów, abyś był świadomy braku jednej ogólnie obowiązującej metody. Badacze częściej niż kiedyś podają pewną miarę skośności, aby pomóc swoim klientom lepiej zrozumieć naturę danych. Przedstawienie miary skośności jest znacznie bardziej skuteczne niż drukowanie wykresu w gazecie i liczenie na to, że czytelnik zdecyduje, jak bardzo różni się ten rozkład od normalnego. Ta różnica może wpływać na wszystko — od znaczenia współczynników korelacji do prawidłowości testów wnioskowania na podstawie zadanych danych.

Na przykład jedna z miar skośności zaproponowana przez Karla Pearsona (tego od współczynnika korelacji Pearsona) jest pokazana tutaj:

$$\text{Skośność} = (\text{Średnia} - \text{Moda}) / \text{Odchylenie standardowe}$$

Jednak bardziej typowym sposobem jest użycie sumy podniesionych do sześciannu wartości standaryzowanych (wartości z) w rozkładzie. Jedna z takich metod obliczania skośności jest następująca:

$$\sum_{i=1}^N z_i^3 / N$$

Jest to po prostu średnia z sześciannów wartości standaryzowanych.

Excel używa odmiany takiej formuły w funkcji SKOŚNOŚĆ ():

$$N \sum_{i=1}^N z_i^3 / ((N-1)(N-2))$$

Po odrobinie zastanowienia można zauważyć, że funkcja Excela zawsze zwraca większą wartość niż prosta średnia sześciannów wartości standaryzowanych. Jeżeli liczba wartości w rozkładzie jest duża, te dwa podejścia są prawie równoważne. Jednak w przypadku tylko pięciu wartości funkcja SKOŚNOŚĆ () zwraca wartość ponad dwukrotnie większą niż średnia sześciannów wartości z . Spójrz na rysunek 7.2, gdzie oryginalne wartości w kolumnie A są po prostu powielone w kolumnie E. Zauważ, że zarówno średnia sześciannów wartości z , jak i wartość zwracana przez funkcję SKOŚNOŚĆ () zależą od liczby danych.

Kurtoza

Rozkład może być symetryczny, ale nadal daleki od normalnego wzorca, gdy jest bardziej wysmukły lub bardziej płaski niż prawdziwa krzywa normalna. Ta cecha jest nazywaną **kurtozą** krzywej.

Rysunek 7.2.

Wraz ze wzrostem liczby wartości różnicę pomiędzy średnią sześciątów wartości z a wartością zwracaną przez funkcję SKOŚNOŚĆ() jest coraz mniejsza

	A	B	C	D	E	F	G
	Wartości oryginalne	Wartości standaryzowane (wartości z)	z^3		Wartości oryginalne	Wartości standaryzowane (wartości z)	z^3
1							
2	2	-0,610257153	-0,22727		2	-0,659153062	-0,28639
3	2	-0,610257153	-0,22727		2	-0,659153062	-0,28639
4	3	-0,271225401	-0,01995		3	-0,292956916	-0,02514
5	3	-0,271225401	-0,01995		3	-0,292956916	-0,02514
6	9	1,762965109	5,479377		9	1,904219956	6,904804
7					2	-0,659153062	-0,28639
8		Średnia z^3 :	0,996987		2	-0,659153062	-0,28639
9		=SKOŚNOŚĆ(A2:A6)	2,077057		3	-0,292956916	-0,02514
10					3	-0,292956916	-0,02514
11					9	1,904219956	6,904804
12					2	-0,659153062	-0,28639
13					2	-0,659153062	-0,28639
14					3	-0,292956916	-0,02514
15					3	-0,292956916	-0,02514
16					9	1,904219956	6,904804
17							
18						Średnia z^3 :	1,256347
19						=SKOŚNOŚĆ(E2:E16)	1,553177

Typy kurtozy

Kilka przymiotników, które opisują naturę kurtozy krzywej, pojawia się prawie wyłącznie w podręcznikach statystyki:

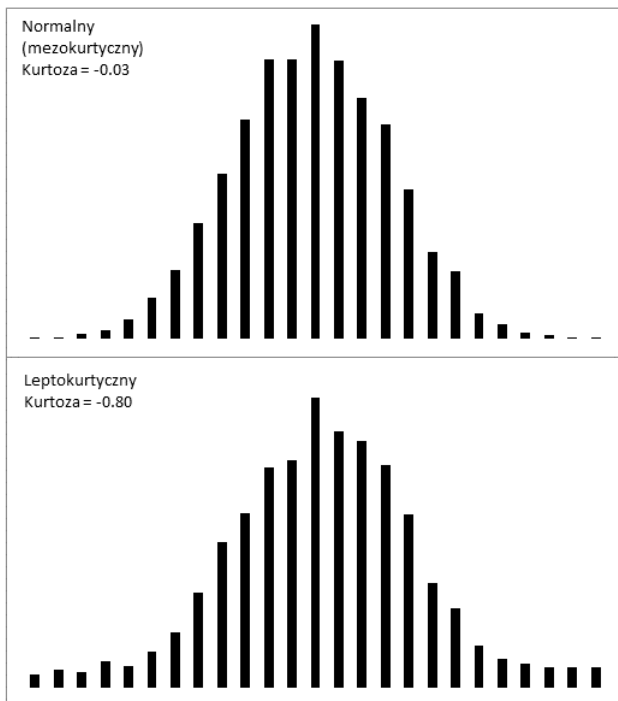
- **Platykurtyczna** krzywa jest bardziej płaska i szeroka niż krzywa normalna.
- **Mezokurtyczna** krzywa ma przeciętną kurtozę. Krzywa normalna jest mezkurtyczna.
- **Leptokurtyczna** krzywa jest bardziej wypukła niż krzywa normalna — pole środkowe jest bardziej wysmukłe. Oznacza to, że większy obszar pod krzywą znajduje się na brzegach. Innymi słowy, grubsze ogony rozkładu zabierają więcej pola ze środka krzywej.

Rozkład t-Studenta (patrz rozdział 8.) jest leptokurtyczny, ale im więcej obserwacji w próbie losowej, tym bardziej przypomina krzywą normalną. Ponieważ większy obszar znajduje się w ogonach rozkładu t-Studenta, konieczne są specjalne porównania w przypadku wykorzystania tego rozkładu do testu średnich dla względnie małej próby losowej. Rozdziały 8. i 9. zawierają dość szczegółowy opis tego problemu, ale zobaczysz, że leptokurtyczny rozkład t-Studenta ma także zastosowanie w analizie regresji (patrz rozdział 12.).

Rysunek 7.3 przedstawia krzywą zbliżoną do normalnej — w dowolnej proporcji, z kurtozą $-0,03$ bardzo bliską zera. Pokazuje także nieco leptokurtyczną krzywą z kurtozą równą $-0,80$.

Rysunek 7.3.

Obserwacje położone bliżej środka krzywej normalnej przesuwają się w kierunku ogonów krzywej leptokurtycznej



Zauważ, że większy obszar pod krzywą leptokurtyczną znajduje się w ogonie rozkładu, a mniejszy w jego środku. Rozkład t-Studenta jest zgodny z tym wzorcem, a testy dotyczące takich statystyk jak średnie biorą to pod uwagę, gdy na przykład odchylenie standardowe jest nieznane, a liczebność próby mała. Gdy większa część pola powierzchni leży w ogonach rozkładu, krytyczne wartości potrzebne do odrzucenia hipotezy zerowej są większe niż w przypadku rozkładu normalnego. Ten efekt znajduje także zastosowanie w konstrukcji przedziałów ufności (opisanych dalej w tym rozdziale).

Obliczanie kurtozy

Przesłanki do obliczania kurtozy są takie same jak w przypadku obliczania skośności — liczba jest często bardziej skutecznym sposobem opisu niż wykres. Co więcej, znajomość oddalenia rozkładu od krzywej normalnej pomaga odbiorcom badań poznać kontekst pozostałych wniosków.

Excel oferuje funkcję arkusza `KURTOZA()` do obliczania kurtozy w zbiorze liczb. Niestety, podobnie jak w przypadku skośności, nie ustalono jednego wzoru na kurtozę. Statystycy zgadzają się jednak co do tego, że zalecane metody zwykle korzystają z pewnej odmiany not standardowych podniesionych do czwartej potęgi.

Oto podręcznikowa definicja kurtozy:

$$\frac{\sum_{i=1}^N z_i^4}{N} - 3$$

W tej definicji N jest liczbą wartości w rozkładzie, a z reprezentuje odpowiednie wartości standaryzowane — jest to każda wartość pomniejszona o średnią i podzielona przez odchylenie standardowe.

Liczba 3 jest odejmowana, aby dla krzywej normalnej otrzymać wynik równy 0. Wtedy dodatnie wartości kurtozy wskazują rozkład leptokurtyczny, a ujemne — rozkład platykurtyczny. Ponieważ wartości standaryzowane są podnoszone do parzystej potęgi, ich suma (a dlatego również średnia) nie może być ujemna. Odejmowanie liczby 3 jest wygodnym sposobem nadania platykurtycznym krzywom ujemnej wartości kurtozy. Niektóre wersje tej formuły nie odejmują tej liczby i zwracają wartość 3 dla krzywej normalnej.

Funkcja `KURTOZA()` jest obliczana według poniższego wzoru, zgodnie z podejściem, które ma na celu korekcję obciążenia w estymacji parametru populacji za pomocą próby:

$$Kurtoza = \frac{N(N+1)}{(N-1)(N-2)(N-3)} \sum_{i=1}^N z_i^4 - \frac{3(N-1)^2}{(N-2)(N-3)}$$

Standaryzowany rozkład normalny

Jedna szczególna wersja normalnego rozkładu ma specjalną wagę. Mowa tu o rozkładzie **normalnym standaryzowanym** lub **normalnym standardowym**. Jego kształt jest taki sam jak dowolnego rozkładu normalnego, ale średnia wynosi 0, a odchylenie standardowe 1. Ta lokalizacja (średnia 0) i rozproszenie (standardowe odchylenie równe 1) sprawiają, że jest on standardowy, co jest bardzo wygodne.

Z powodu tych dwóch charakterystyk natychmiast znasz skumulowane pole powierzchni poniżej dowolnej wartości. W jednostkowym rozkładzie normalnym wartość 1 jest oddalona w prawo o jedno standardowe odchylenie od średniej 0 i dlatego 84% pola powierzchni przypada po jej lewej stronie. Wartość -2 znajduje się dwa odchylenia standardowe poniżej średniej równej 0, więc 2,275% pola powierzchni leży po jej lewej stronie.

Z drugiej strony założmy, że pracujesz z rozkładem, który ma średnią 7,63 centymetra i odchylenie standardowe 0,124 centymetra — przypuśćmy, że reprezentuje średnicę części maszyny, której rozmiar musi być precyzyjny. Jeżeli ktoś powie Ci, że jedna z części maszyny ma średnicę równą 7,816, prawdopodobnie będziesz musiał myśleć przez chwilę, zanim stwierdzisz, że jest to półtora odchylenia standardowego powyżej średniej. Ale jeżeli używasz standaryzowanego rozkładu normalnego jako miary, po usłyszeniu wyniku 1,5 będziesz znał dokładne położenie wymiaru tej części maszyny w rozkładzie.

Interpretacja znaczenia wartości jest szybsza i łatwiejsza, jeżeli używasz standaryzowanego rozkładu normalnego. Excel ma funkcje arkusza przeznaczone do rozkładu normalnego i są one łatwe w użyciu. Excel udostępnia także funkcje skrojone specjalnie do jednostkowego rozkładu normalnego, które są jeszcze łatwiejsze w użyciu — nie trzeba podawać średniej ani odchylenia standardowego rozkładu, ponieważ są one znane. Następny podrozdział skupia się na tych funkcjach zarówno w Excelu 2010, jak i jego wcześniejszych wersjach.

Funkcje Excela dla rozkładu normalnego

Excel nazywa funkcje, które dotyczą rozkładu normalnego, w taki sposób, abyś zawsze wiedział, czy masz do czynienia z dowolnym rozkładem normalnym, czy ze standaryzowanym rozkładem normalnym ze średnią 0 i odchyleniem standardowym równym 1.

Excel odwołuje się do standaryzowanego rozkładu normalnego jako „standardowego” normalnego i dlatego używa litery S w nazwie funkcji. Dlatego funkcja `ROZKŁ.NORMALNY()` dotyczy dowolnego rozkładu normalnego, podczas gdy funkcja zgodności `ROZKŁAD.NORMALNY.S()` i funkcja spójności `ROZKŁ.NORMALNY.S()` odnoszą się do standaryzowanego rozkładu normalnego.

Funkcja `ROZKŁ.NORMALNY()`

Załóżmy, że interesuje Cię rozkład w populacji poziomów lipoprotein o wysokiej gęstości (ang. *high-density lipoprotein*, HDL), czyli tzw. „dobrego cholesterolu”, wśród osób dorosłych powyżej 20. roku życia. Ta zmienna jest normalnie mierzona w miligramach na decylitr krwi (mg/dl). Zakładając, że poziomy HDL mają rozkład normalny (a mają), możesz dowiedzieć się więcej na temat rozkładu HDL w populacji, stosując wiedzę na temat krzywej normalnej. Sposobem, żeby to zrobić, jest użycie funkcji Excela `ROZKŁ.NORMALNY()`.

Składnia funkcji `ROZKŁ.NORMALNY()`

Funkcja `ROZKŁ.NORMALNY()` przyjmuje następujące dane jako argumenty:

- **x** — to wartość w rozkładzie, którą obliczasz. Jeżeli obliczasz poziomy HDL, możesz być zainteresowany konkretnym poziomem — powiedzmy, 60. Ta konkretna wartość jest tą, którą podasz jako pierwszy argument funkcji `ROZKŁ.NORMALNY()`.
- **Średnia** — drugim argumentem jest średnia rozkładu, nad którym pracujesz. Załóżmy, że średni poziom HDL wśród ludzi powyżej 20. roku życia wynosi 54,3.
- **Odchylenie standardowe** — trzecim argumentem jest odchylenie standardowe rozkładu, nad którym pracujesz. Załóżmy, że standardowe odchylenie poziomów HDL wynosi 15.

- **Skumulowany** — czwarty argument wskazuje, czy chcesz otrzymać skumulowane prawdopodobieństwo poziomów HDL od 0 do x (równego w tym przykładzie 60), czy prawdopodobieństwo odpowiadające poziomowi HDL równemu dokładnie x (czyli 60). Jeżeli chcesz otrzymać skumulowane prawdopodobieństwo, podaj **PRAWDA** jako czwarty argument. W przeciwnym przypadku użyj wartości **FALSZ**.

Żądanie skumulowanego prawdopodobieństwa

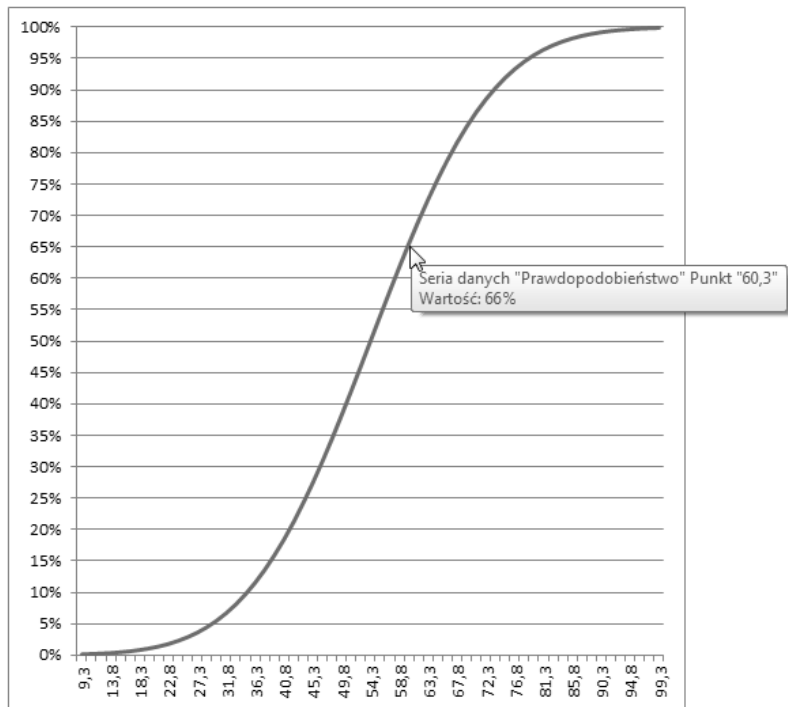
Formuła:

`=ROZKŁ.NORMALNY(60;54,3;15;PRAWDA)`

zwraca 0,648, czyli 64,8%. Oznacza to, że 64,8% pola powierzchni pod rozkładem poziomów HDL znajduje się pomiędzy 0 a 60 mg/dl. Ten wynik został pokazany na rysunku 7.4.

Rysunek 7.4.

Możesz dostosować liczbę linii siatki, formatując oś pionową, aby pokazywała więcej lub mniej głównych jednostek

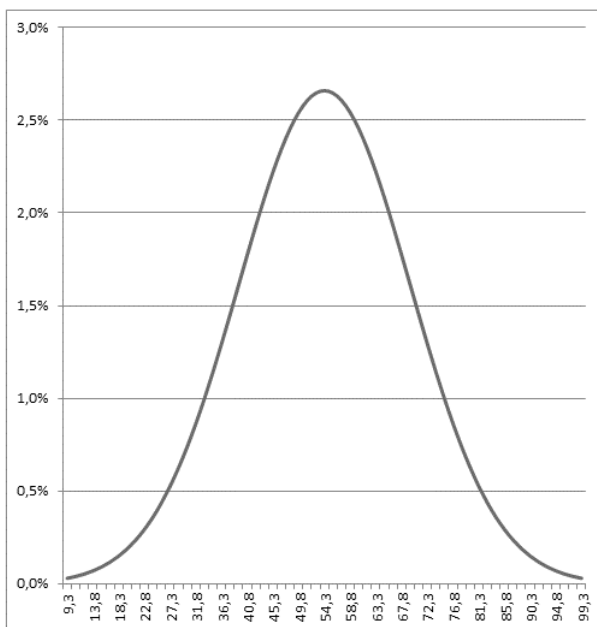


Jeżeli przytrzymasz wskaźnik myszy nad linią, która pokazuje skumulowane prawdopodobieństwo, zobaczysz małe okno podręczne informujące, który punkt danych wskazujesz i jakie jest jego położenie na osiach poziomej i pionowej. Raz utworzony wykres może poinformować Cię o prawdopodobieństwie związanym z dowolnym wykreślonym punktem, nie tylko opisanym tutaj 60 mg/dl. Jak widać na rysunku 7.4, możesz użyć albo linii siatki wykresu, albo wskaźnika myszy do wyznaczenia, że miara na przykład 60,3 mg/dl lub niższa jest zaliczana do około 66% populacji.

Żądanie oszacowania punktu

Inaczej wygląda to, gdy jako czwarty (*skumulowany*) argument funkcji `ROZKŁ.NORMALNY()` wybierzesz `FAŁSZ`. W tym przypadku funkcja zwraca prawdopodobieństwo związane z konkretnym punktem określonym w pierwszym argumencie. Stosuj wartość `FAŁSZ` argumentu *skumulowany*, jeżeli chcesz znać wysokość krzywej normalnej dla konkretnej wartości obliczanego rozkładu. Rysunek 7.5 przedstawia jeden ze sposobów użycia funkcji `ROZKŁ.NORMALNY()` z argumentem *skumulowany* równym `FAŁSZ`.

Rysunek 7.5.
Wysokość krzywej
w dowolnym punkcie to
wartość funkcji gęstości
prawdopodobieństwa



Nieczęsto zdarza się, że trzeba wyliczyć konkretną wysokość krzywej normalnej dla konkretnej wartości, ale jeżeli tak się stanie — na przykład po to, by narysować krzywą, która pomoże zobrazować wyniki — przydaje się podanie argumentu *skumulowany* równego `FAŁSZ`. (Ta wartość — prawdopodobieństwo odpowiadające konkretnemu punktowi albo, inaczej, wysokość krzywej w tym punkcie — jest określana jako **funkcja gęstości prawdopodobieństwa**).

Jeżeli używasz wersji Excela wcześniejszej niż 2010, możesz użyć funkcji zgodności `ROZKŁAD.NORMALNY()`. Jest ona taka sama jak `ROZKŁ.NORMALNY()` pod względem argumentów i zwracanych wartości.

Funkcja ROZKŁ.NORMALNY.ODWR()

Ze względów praktycznych zwykle funkcja ROZKŁ.NORMALNY() jest potrzebna po fakcie, czyli wtedy, gdy dane są już zebrane i znasz średnią oraz odchylenie standardowe próby losowej lub populacji. Rodzi się wtedy pytanie: gdzie dana wartość znajduje się w rozkładzie normalnym? Ta wartość może być średnią próby losowej, którą chcesz porównać z populacją, lub pojedynczą obserwacją, którą chcesz umieścić w kontekście większej grupy.

W tym przypadku możesz przekazać te informacje do funkcji ROZKŁ.NORMALNY(), która określi prawdopodobieństwo obserwowania wartości mniejszej lub równej zadanej (*skumulowany* = PRAWDA) bądź wartość funkcji gęstości (*skumulowany* = FAŁSZ). Możesz następnie porównać to prawdopodobieństwo z przyjętym w eksperymencie współczynnikiem α .

Funkcja ROZKŁ.NORMALNY.ODWR() jest blisko związana z funkcją ROZKŁ.NORMALNY() i pozwala spojrzeć na zagadnienie z trochę innej perspektywy. Zamiast zwracać wartość przedstawiającą pole powierzchni — czyli prawdopodobieństwo — funkcja ROZKŁ.NORMALNY.ODWR() zwraca wartość, która reprezentuje punkt na osi poziomej krzywej rozkładu normalnego. Jest to punkt, który podaje się jako pierwszy argument funkcji ROZKŁ.NORMALNY().

Na przykład w poprzedniej sekcji pokazaliśmy, że formuła:

=ROZKŁ.NORMALNY(60;54,3;15;PRAWDA)

zwraca wartość 0,648. Wartość 60 jest przynajmniej tak wielka jak 64,8% obserwacji w rozkładzie normalnym o średniej 54,3 i odchyleniu standardowym równym 15.

Z drugiej strony formuła:

=ROZKŁ.NORMALNY.ODWR(0,648;54,3;15)

zwraca wartość 60. Jeżeli rozkład ma średnią 54,3 i odchylenie standardowe równe 15, wtedy 64,8% obserwacji ma wartość 60 lub mniejszą. Ten przykład jest dobrą ilustracją. Zwykle nie zwróciłbyś uwagi, że 64,8% obserwacji leży poniżej konkretnej wartości.

Załóżmy jednak, że podczas przygotowywania projektu badawczego zdecydowałeś, że uznasz za wiarygodny efekt pewnego oddziaływania (np. kuracji medycznej) tylko wtedy, gdy średnia grupy eksperymentalnej będzie się znajdowała w górnych 5% populacji. (Jest to spójne z tradycyjnym podejściem do badań z wykorzystaniem hipotezy zerowej, które zostanie znacznie bardziej szczegółowo opisane w rozdziałach 8. i 9.). W tym przypadku będziesz chciał wiedzieć, jaki wynik definiuje górne 5%.

Jeżeli znasz średnią i odchylenie standardowe, funkcja ROZKŁ.NORMALNY.ODWR() wykona to zadanie za Ciebie. Zajmijmy się dalej populacją ze średnią 54,3 i odchyleniem standardowym równym 15. Formuła:

=ROZKŁ.NORMALNY.ODWR(0,95;54,3;15)

zwraca 78,97. Pięć procent obserwacji o rozkładzie normalnym, który ma średnią równą 54,3 i odchylenie standardowe równe 15, leży powyżej wartości 78,97.

Jak widzisz, w formule zastosowałem 0,95 jako pierwszy argument funkcji `ROZKŁ.NORMALNY.ODWR()`. Jest tak, ponieważ funkcja `ROZKŁ.NORMALNY.ODWR()` zakłada skumulowane prawdopodobieństwo — zauważ, że w przeciwieństwie do funkcji `ROZKŁ.NORMALNY()`, funkcja `ROZKŁ.NORMALNY.ODWR()` nie ma czwartego argumentu skumulowany. Dlatego pytanie, jaka wartość odcina górne 5% rozkładu, jest równoważne z pytaniem, jaka wartość odcina dolne 95% rozkładu.

W tym kontekście wybór użycia funkcji `ROZKŁ.NORMALNY()` lub `ROZKŁ.NORMALNY.ODWR()` zależy głównie od rodzaju poszukiwanych informacji. Jeżeli chcesz wiedzieć, jakie jest prawdopodobieństwo, że zaobserwujesz liczbę tak dużą jak X , podaj tę wartość funkcji `ROZKŁ.NORMALNY()`, aby otrzymać prawdopodobieństwo. Jeżeli chcesz poznać liczbę, która służy jako ograniczenie pola powierzchni — pola, które odpowiada zadanemu prawdopodobieństwu — podaj pole powierzchni do funkcji `ROZKŁ.NORMALNY.ODWR()`, aby otrzymać tę liczbę.

W każdym z tych przypadków musisz podać średnią i odchylenie standardowe. W przypadku funkcji `ROZKŁ.NORMALNY` potrzebny jest dodatkowy argument określający, czy interesuje Cię skumulowane prawdopodobieństwo, czy wartość funkcji gęstości w punkcie.

Funkcja spójności `ROZKŁ.NORMALNY.ODWR()` nie jest dostępna w wersjach Excela wcześniejszych niż 2010, ale zamiast niej możesz użyć funkcji zgodności `ROZKŁAD.NORMALNY.ODW()`. Argumenty i wyniki są takie same jak w przypadku funkcji `ROZKŁ.NORMALNY.ODWR()`.

Korzystanie z funkcji `ROZKŁ.NORMALNY.S()`

Można by wiele pisać o wyrażaniu odległości, wag, czasu itp. w ich oryginalnych jednostkach pomiarowych. Służy do tego funkcja `ROZKŁ.NORMALNY()`. Kiedy jednak chcesz używać standardowej jednostki miary do zmiennej o rozkładzie normalnym, powinieneś pomyśleć o funkcji `ROZKŁ.NORMALNY.S()`. Litera *S* w nazwie funkcji oznacza oczywiście **standardowy**.

Stosowanie funkcji `ROZKŁ.NORMALNY.S()` jest szybsze, ponieważ nie trzeba podawać średniej ani odchylenia standardowego. Średnia (0) i odchylenie standardowe (1) standaryzowanego rozkładu normalnego są znane z definicji. Wszystko, czego potrzebuje funkcja `ROZKŁ.NORMALNY.S()`, to wartość standaryzowana (wartość z) oraz wskazanie, czy ma obliczyć skumulowane pole powierzchni (**PRAWDA**), czy wartość funkcji gęstości w punkcie (**FALSZ**). Funkcja korzysta z prostej składni:

```
=ROZKŁ.NORMALNY.S(z;skumulowany)
```

Dlatego formuła:

```
=ROZKŁ.NORMALNY.S(1,5;PRAWDA)
```

informuje Cię, że 93,3% pola powierzchni pod krzywą znajduje się po lewej stronie wartości standaryzowanej równej 1,5. W rozdziale 3., „Rozrzut — jak się rozpraszają wartości”, znajdziesz wprowadzenie do koncepcji wartości standaryzowanych, czyli wartości z .

OSTRZEŻENIE

Funkcja zgodności ROZKŁAD.NORMALNY.S() jest dostępna w wersjach Excela wcześniejszych niż 2010. Jest to jedyna z funkcji zgodności rozkładu normalnego, której lista argumentów jest inna niż związanej z nią funkcji spójności. ROZKŁAD.NORMALNY.S() nie ma argumentu *skumulowany* — zwraca domyślnie skumulowane pole powierzchni po lewej stronie argumentu z. W Excelu pojawi się ostrzeżenie o błędzie, jeżeli podasz argument skumulowany do funkcji ROZKŁAD.NORMALNY.S(). Jeżeli chcesz obliczyć wartość punktową zamiast skumulowanego prawdopodobieństwa, powinieneś użyć funkcji ROZKŁAD.NORMALNY(), podając 0 jako drugi argument, a 1 jako trzeci. Te dwa argumenty określają wspólnie jednostkowy rozkład normalny. Następnie możesz podać wartość FAŁSZ jako czwarty argument funkcji ROZKŁAD.NORMALNY(). Oto przykład:

=ROZKŁAD.NORMALNY(1;0;1;FAŁSZ)

Korzystanie z funkcji ROZKŁ.NORMALNY.S.ODWR()

Jeszcze łatwiej jest użyć funkcji odwrotnej do ROZKŁ.NORMALNY.S(), którą jest ROZKŁ.NORMALNY.S.ODWR(). Jedyny przyjmowany przez tę funkcję argument to prawdopodobieństwo:

=ROZKŁ.NORMALNY.S.ODWR(0,95)

Ta formuła zwraca 1,64, co oznacza, że 95% pola powierzchni pod krzywą normalną leży po lewej stronie wartości z równej 1,64. Jeżeli uczęszczałeś na kurs podstaw wnioskowania statystycznego, ta liczba prawdopodobnie wygląda znajomo, tak znajomo jak 1,96, która odcina 97,5% rozkładu.

Te liczby występują powszechnie, ponieważ są związane z pojawiającymi się regularnie pod tabelami w prasowych i internetowych raportach wpisami „ $p < 0,05$ ” i „ $p < 0,025$ ” — koleinami, w które nie chciałbyś wpaść. Znacznie więcej informacji na temat takich wpisów znajduje się w rozdziałach 8. i 9. w kontekście rozkładu t-Studenta (który jest blisko związany z rozkładem normalnym).

Funkcja zgodności ROZKŁAD.NORMALNY.S.ODW() przyjmuje ten sam argument i zwraca te same wyniki co ROZKŁ.NORMALNY.S.ODWR().

Jest jeszcze jedna funkcja arkusza Excela, która dotyczy bezpośrednio rozkładu normalnego — UFNOŚĆ.NORM(). Przed właściwym omówieniem jej celu i zastosowania niezbędne jest najpierw poznanie odrobiny podstaw teoretycznych.

Przedziały ufności i rozkład normalny

Przedział ufności to rozstęp wartości, który daje użytkownikowi poczucie, jak dokładnie dana statystyka szacuje parametr. Najbardziej znanym zastosowaniem przedziału ufności jest prawdopodobnie „margines błędu” umieszczany w wiadomościach na temat sondaży:

„Margines błędu wynosi ± 3 punkty procentowe”. Jednak przedziały ufności są użyteczne w kontekście, który znacznie wykracza poza tę prostą sytuację.

Przedziały ufności mogą być używane z rozkładami, które nie są normalne — takimi, które są mocno skośne lub w inny sposób odróżniają się od normalnych. Jednak najprościej je poznać na przykładzie rozkładów symetrycznych, dlatego właśnie tym tematem się teraz zajmujemy. Nie myśl jednak, że możesz używać przedziałów ufności tylko do rozkładów normalnych.

Znaczenie przedziału ufności

Żałujemy, że zmierzyles poziom HDL w krwi 100 osób dorosłych na specjalnej diecie i obliczyłeś średnią 50 mg/dl z odchyleniem standardowym równym 20. Jesteś świadomy, że średnia jest statystyką, a nie parametrem populacji, i że inna próba 100 dorosłych na tej samej diecie prawdopodobnie zwróci inną wartość średniej. Po wielu powtórzonych losowaniach prób średnia całkowita — czyli średnia średnich prób losowych — okaże się bardzo, bardzo bliska parametru populacji.

Jednak Twoje zasoby nie są tak duże i masz zamiar dokonać analizy tylko z tą jedną statystyką, równą 50 mg/dl, którą obliczyłeś dla swojej próby. Chociaż wartość 20, którą obliczyłeś dla odchylenia standardowego, jest statystyką, jest taka sama jak znane odchylenie standardowe populacji równe 20. Możesz teraz użyć standardowego odchylenia próby i liczby stabilizowanych wartości HDL w celu wyznaczenia sensu w estymacji za pomocą tej próby.

Robisz to, konstruując przedział ufności wokół średniej równej 50 mg/dl. Przypuśćmy, że przedział rozciąga się od 45 do 55. (I tutaj możesz zobaczyć związek z „ ± 3 punkty procentowe”). Czy to oznacza, że prawdziwa średnia populacji znajduje się gdzieś pomiędzy 45 a 55?

Nie, chociaż może tak być. Tak jak istnieje wiele możliwych prób, które mógłbyś wylosować, ale tego nie zrobiłeś, istnieje wiele możliwych poziomów ufności, które mógłbyś wtedy skonstruować wokół średnich tych prób. Jak zobaczysz, poziom ufności konstruuje się w taki sposób, że gdybyś wziął dużo więcej średnich i umieścił wokół nich poziomy ufności, 95% poziomów ufności objęłoby prawdziwą średnią populacji. Dla konkretnego poziomu ufności, który skonstruowałeś, prawdopodobieństwo, że prawdziwa średnia populacji mieści się w tym przedziale, wynosi albo 1, albo 0 — przedział pokrywa średnią albo nie.

Jednak bardziej racjonalne jest założenie, że przedział ufności, który przyjąłeś, jest jednym z 95%, które zawierają średnią populacji, niż założenie sytuacji odwrotnej. Dlatego będziesz skłonny wierzyć z 95-procentową ufnością, że ten przedział jest jednym z tych, które zawierają średnią populacji.

Chociaż pisałem w tym punkcie o 95-procentowych przedziałach ufności, możesz skonstruować także 90- lub 99-procentowe przedziały ufności czy nawet o dowolnym innym stopniu ufności, który ma dla Ciebie sens w danej sytuacji. Niebawem przekonasz się, jak

Twoje wybory podczas konstruowania przedziału wpływają na samą jego naturę. Opis stanie się prostszy, jeżeli na chwilę zapomnisz o swym niedowierzaniu. W skrócie: zamierzam poprosić Cię o wyobrażenie sobie sytuacji, w której znasz odchylenie standardowe miary w populacji, ale nie wiesz, jaka jest średnia tej populacji. To sytuacja nietypowa, ale jak najbardziej możliwa.

Konstruowanie przedziału ufności

Przedział ufności dla średniej, zgodnie z opisem w poprzednim punkcie, wymaga następujących elementów składowych:

- samej średniej,
- odchylenia standardowego obserwacji,
- liczby obserwacji w próbie,
- poziomu ufności, który chcesz zastosować do przedziału ufności.

Zaczynając od poziomu ufności, założmy, że chcesz utworzyć 95-procentowy przedział ufności. Chcesz skonstruować go w taki sposób, żeby średnio w 95 przypadkach na 100 pokrył prawdziwą średnią populacji.

Ponieważ w tym przypadku masz do czynienia z rozkładem normalnym, możesz wpisać następujące formuły do arkusza:

=ROZKŁ.NORMALNY.S.ODWR(0,025)

=ROZKŁ.NORMALNY.S.ODWR(0,975)

Funkcja ROZKŁ.NORMALNY.S.ODWR() opisana w poprzednim podrozdziale zwraca wartość standaryzowaną, którą ma po lewej stronie ułamek pola powierzchni pod krzywą zadany jako argument. Dlatego funkcja ROZKŁ.NORMALNY.S.ODWR(0,025) zwraca -1,96. Jest to wartość z, po której lewej stronie leży 0,025, czyli 2,5% pola pod krzywą.

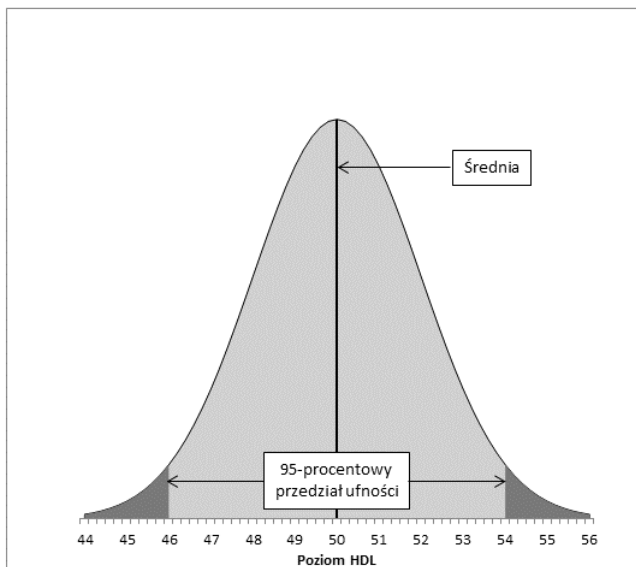
Podobnie funkcja ROZKŁ.NORMALNY.S.ODWR(0,975) zwraca 1,96, czyli wartość z, która ma 97,5% pola pod krzywą po lewej stronie. Inaczej można powiedzieć, że 2,5% pola pod krzywą leży po jej prawej stronie. Te liczby są pokazane na rysunku 7.6.

Pole powierzchni pod krzywą na rysunku 7.6 i pomiędzy wartościami 46,1 a 53,9 na osi poziomej obejmuje 95% pola powierzchni pod krzywą. Krzywa w teorii rozszerza się nieskończenie w lewo i w prawo, więc wszystkie możliwe wartości średnich populacji są uwzględnione. 95% możliwych wartości leży wewnątrz 95-procentowego przedziału ufności pomiędzy 46,1 a 53,9.

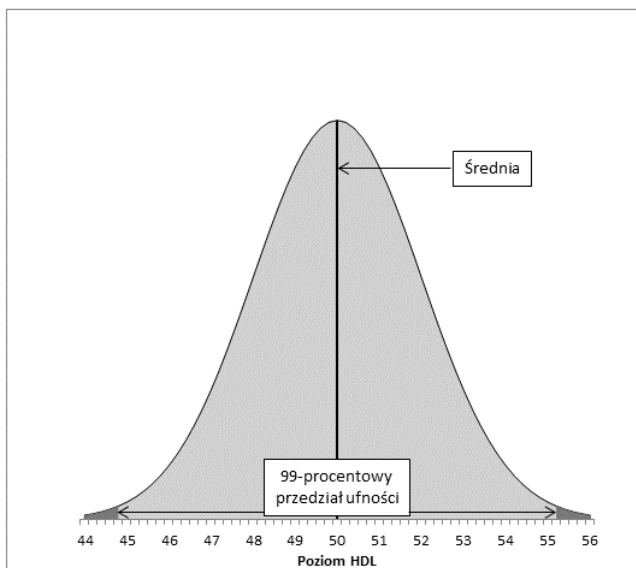
Liczby 46,1 i 53,9 zostały wybrane tak, aby obejmowały 95%. Jeżeli chciałbyś otrzymać 99-procentowy przedział ufności (lub jakiś inny, z mniejszym lub większym prawdopodobieństwem stanowiący jeden z przedziałów obejmujących średnią populacji), wybrałbyś inne liczby. Rysunek 7.7 przedstawia 99-procentowy przedział ufności wokół średniej próby równej 50.

Rysunek 7.6.

Dostosowanie ograniczenia wartości z reguluje poziom ufności. Porównaj rysunki 7.6 i 7.7

**Rysunek 7.7.**

Rozszerzanie przedziału daje większą pewność, że pokryje parametr populacji, ale nieuchronnie skutkuje mniej dokładnym oszacowaniem



Na rysunku 7.7 99-procentowy przedział ufności rozszerza się z 44,8 do 55,2, czyli zwiększa się o 2,6 jednostki w stosunku do 95-procentowego przedziału ufności przedstawionego na rysunku 7.6. Na sto 99-procentowych przedziałów ufności skonstruowanych wokół średnich 100 prób średnio 99 (nie 95, jak było poprzednio) pokryłoby średnią populacji. Dodatkowa ufność jest zapewniana przez rozszerzenie przedziału. Ta wymiana zawsze dotyczy przedziałów ufności. Im większy przedział, tym bardziej precyzyjnie nakreślone

jego granice, ale mniej takich przedziałów pokryje poszukiwaną statystykę (tutaj średnią). Im szerszy przedział, tym mniej precyzyjnie określasz granice, ale większa liczba przedziałów pokryje daną statystykę.

Poza określeniem poziomu ufności jedynym czynnikiem, który jest pod Twoją kontrolą, jest liczebność próby losowej. Nie możesz nakazać, aby odchylenie standardowe było mniejsze, ale możesz dobrać liczniejsze próby. Jak zobaczysz w rozdziałach 8. i 9., odchylenie standardowe użyte w przedziałach ufności wokół średniej próby nie jest odchyleniem standardowym poszczególnych surowych wyników. Jest to odchylenie standardowe podzielone przez pierwiastek kwadratowy liczebności próby i nosi nazwę **błędu standardowego średniej arytmetycznej**.

Zbiór danych używany do kreślenia wykresów na rysunkach 7.6 i 7.7 ma odchylenie standardowe równe 20, które jest uważane za takie samo jak odchylenie standardowe populacji. Liczebność próby wynosi 100. Dlatego błąd standardowy średniej wynosi:

$$\text{Błąd standardowy} = \frac{20}{\sqrt{100}}$$

czyli 2.

W celu ukończenia konstrukcji przedziału ufności należy pomnożyć błąd standardowy średniej przez wartości standaryzowane, które odcinają interesujący Cię poziom ufności. Rysunek 7.6 przedstawia na przykład 95-procentowy przedział ufności. Interwał musi być zbudowany tak, aby 95% leżało poniżej krzywej i wewnątrz przedziału — dlatego 5% musi leżeć poza przedziałem, podzielone po równo 2,5% pomiędzy dwa ogony.

Oto miejsce, gdzie wchodzi do gry funkcja ROZKŁ. NORMALNY. S. ODWR (). Wcześniej w tej sekcji zostały użyte dwie formuły:

$$\begin{aligned} &= \text{ROZKŁ. NORMALNY. S. ODWR}(0,025) \\ &= \text{ROZKŁ. NORMALNY. S. ODWR}(0,975) \end{aligned}$$

Zwracają one wartości standaryzowane $-1,96$ i $1,96$, które stanowią granice odpowiednio dla 2,5% i 97,5% standaryzowanego rozkładu normalnego. Jeżeli pomnożysz każdą z nich przez błąd standardowy równy 2 i dodasz średnią równą 50, otrzymasz 46,1 i 53,9, granice 95-procentowego przedziału ufności o średniej 50 i błędzie standardowym równym 2.

Jeżeli chcesz mieć 99-procentowy poziom ufności, zastosuj formuły:

$$\begin{aligned} &= \text{ROZKŁ. NORMALNY. S. ODWR}(0,005) \\ &= \text{ROZKŁ. NORMALNY. S. ODWR}(0,995) \end{aligned}$$

aby otrzymać wyniki $-2,58$ i $2,58$. Te noty standardowe odcinają połowę jednego procenta ze standaryzowanego rozkładu normalnego na każdym końcu. Pozostały obszar pod krzywą to 99%. Pomnożenie każdej wartości standaryzowanej przez 2 i dodanie średniej równej 50 da w wyniku 44,8 i 55,2, czyli granice 99-procentowego przedziału ufności dla średniej równej 50 i odchylenia standardowego równego 2.

W tym momencie może pomóc odwrócenie uwagi od arytmetyki i skupienie się zamiast tego na teorii. Dowolna wartość standaryzowana jest pewną liczbą odchyłeń standardowych — dlatego wartość z równa 1,96 odpowiada punktowi, który znajduje się 1,96 odchyłeń standardowych powyżej średniej, a $-1,96$ znajduje się 1,96 odchyłeń standardowych poniżej średniej.

Ponieważ natura krzywej normalnej została przestudiowana tak wyczerpująco, wiemy, że 95% pola powierzchni pod krzywą normalną znajduje się pomiędzy 1,96 odchylenia standardowego poniżej średniej a 1,96 odchylenia standardowego powyżej średniej.

Gdy chcesz umieścić przedział ufności wokół średniej arytmetycznej próby losowej, zaczynasz od zdecydowania, jaki procent średnich innych prób, jeżeli byłyby zebrane i obliczone, chciałbyś znaleźć w tym przedziale. Dlatego jeżeli zdecydujesz, że chciałbyś, aby 95% możliwych średnich prób znajdowało się w Twoim przedziale ufności, umieściłbyś go 1,96 odchylenia standardowego powyżej i poniżej średniej arytmetycznej próby losowej.

Ale jak duże jest odpowiednie odchylenie standardowe? W tej sytuacji odpowiednimi jednostkami są same wartości średnich. Musisz znać odchylenie standardowe nie oryginalnych indywidualnych obserwacji, ale średnich, które są obliczone na ich podstawie. To odchylenie standardowe ma specjalną nazwę — błąd standardowy średniej.

Dzięki obliczeniom matematycznym **oraz** długiemu doświadczeniu w kwestii zachowania liczb wiemy, że dobrym estymatorem standardowego odchylenia wartości średnich jest odchylenie standardowe poszczególnych wyników podzielone przez pierwiastek kwadratowy liczebności próby. Jest to odchylenie standardowe, które chcesz wykorzystać do wyznaczenia przedziału ufności.

W przykładzie analizowanym w tym podrozdziale odchylenie standardowe wynosi 20, a rozmiar próby to 100, dlatego błąd standardowy średniej wynosi 2. Gdy obliczysz 1,96 odchylenia standardowego poniżej i powyżej średniej równej 50, otrzymasz wartości 46,1 i 53,9. Jest to Twój 95-procentowy przedział ufności. Jeżeli weźmiesz inne 99 próbek z tej populacji, około 95 ze 100 podobnych przedziałów ufności pokryje średnią populacji. Założenie, że przedział ufności, który zbudowałeś, jest jednym z 95 na 100, które obejmują średnią populacji, jest rozsądne. Nie byłoby takim stwierdzenie, że jest to jeden z pozostałych 5 na 100, które nie obejmują średniej populacji.

Funkcje arkusza Excela, które wyznaczają przedziały ufności

Rozważania w poprzednim punkcie na temat używania rozkładu normalnego zakładały, że znasz odchylenie standardowe populacji. Nie jest to niewiarygodne założenie, ale jest prawdą, że często nie znasz odchylenia standardowego populacji i musisz je oszacować na podstawie próby. Są dwa różne rozkłady, do których potrzebujesz dostępu w zależności od tego, czy znasz odchylenie standardowe, czy je szacujesz. Jeżeli je znasz, odwołujesz się do rozkładu normalnego. Jeżeli je szacujesz na podstawie próby, używasz rozkładu t-Studenta.

Excel 2010 ma dwie funkcje arkusza: $UFNOŚĆ.NORM()$ i $UFNOŚĆ.T()$, które pomagają obliczyć **szerokość** przedziałów ufności. Funkcję $UFNOŚĆ.NORM()$ stosujesz, gdy znasz odchylenie standardowe populacji dla danej miary (takie jak w przykładzie tego rozdziału dotyczącym poziomów HDL). Funkcję $UFNOŚĆ.T()$ stosujesz, gdy nie znasz odchylenia standardowego populacji i szacujesz je na podstawie danych próby. Więcej informacji na temat tego rozróżnienia i wyboru pomiędzy użyciem rozkładu normalnego a rozkładu t-Studenta znajdziesz w rozdziałach 8. i 9.

Wersje Excela wcześniejsze niż 2010 miały tylko funkcję $UFNOŚĆ()$. Jej argumenty i wyniki są identyczne z tymi funkcji $UFNOŚĆ.NORM()$. Przed wersją 2010 nie było żadnej pojedynczej funkcji arkusza, która by zwracała przedział ufności na podstawie rozkładu t-Studenta. Jednak, jak zobaczysz w tym punkcie, bardzo łatwo można zastąpić funkcję $UFNOŚĆ.T()$ za pomocą funkcji $ROZKŁ.T.ODWR()$ lub $ROZKŁAD.T.ODW()$. Możesz zastąpić funkcję $UFNOŚĆ.NORM()$ za pomocą funkcji $ROZKŁ.NORMALNY.S.ODWR()$ lub $ROZKŁAD.NORMALNY.S.ODW()$.

Korzystanie z funkcji $UFNOŚĆ.NORM()$ i $UFNOŚĆ()$

Rysunek 7.8 przedstawia mały zbiór danych w komórkach A2:A17. Jego średnia znajduje się w komórce B2, a odchylenie standardowe **populacji** w komórce C2.

Rysunek 7.8.

Możesz skonstruować przedział ufności, używając funkcji ufności albo funkcji rozkładu normalnego

G2		fx		=UFNOŚĆ.NORM(F2;C2;ILE.LICZB(A2:A17))					
	A	B	C	D	E	F	G	H	I
		HDL	Średnia HDL	Odchylenie standardowe populacji		Alfa	Półowa szerokości przedziału		
1		88	57,19	22,00		0,05	10,78		
2		64							
3		50			Przedział ufności:		46,41	do	67,97
4		67							
5		45							
6		86							
7		71							
8		68							
9		36							
10		20							
11		57							
12		49							
13		37							
14		94							
15		39							
16		44							
17									

Na rysunku 7.8 wartość nazwana **alfa** (α) znajduje się w komórce F2. Użycie tego terminu jest spójne z użyciem w innych kontekstach, na przykład przy weryfikacji hipotez. Jest to pole powierzchni pod krzywą, które jest poza granicami przedziału ufności. Na rysunku 7.6 α jest sumą zacienionych obszarów w ogonach krzywej. Każdy zacieniony obszar obejmuje 2,5% całkowitego obszaru, więc α wynosi 5% lub 0,05. Wynikiem jest 95-procentowy przedział ufności.

Komórka G2 na rysunku 7.8 pokazuje, jak użyć funkcji `UFNOŚĆ.NORM()`. Zauważ, że możesz w ten sam sposób użyć funkcji zgodności `UFNOŚĆ()`. Oto składnia:

`=UFNOŚĆ.NORM(alfa;odchylenie_standardowe;rozmiar)`

gdzie *rozmiar* dotyczy liczebności próby. Ta funkcja wprowadzona w komórce G2 ma α równe 0,05, odchylenie standardowe populacji równe 22 i 16 wartości w próbie:

`=UFNOŚĆ.NORM(F2;C2;ILE.LICZB(A2:A17))`

Wynikiem funkcji z tymi argumentami jest 10,78. Komórki G4 i I4 prezentują odpowiednio górną i dolną granicę 95-procentowego przedziału ufności.

Warto zwrócić uwagę na kilka punktów:

- Została użyta funkcja `UFNOŚĆ.NORM()`, a nie `UFNOŚĆ.T()`. Jest tak, ponieważ znasz odchylenie standardowe populacji i nie musisz go szacować na podstawie standardowego odchylenia próby. Jeżeli miałbyś estymować wartość populacji na podstawie próby, użyłbyś funkcji `UFNOŚĆ.T()`, zgodnie z opisem w następnym punkcie.
- Ponieważ suma poziomu ufności (np. 95%) i α zawsze równa jest 100%, Microsoft mógł zamiast pytania o α pytać o poziom ufności. Standardem jest odwoływanie się do przedziałów ufności w kategoriach poziomów ufności, takich jak 95%, 90%, 99% itd. Microsoft mógł bardziej uwzględnić potrzeby klientów, wybierając użycie poziomu ufności zamiast α jako pierwszego argumentu funkcji.
- Dokumentacja pomocy programu stwierdza, że funkcja `UFNOŚĆ.NORM()` oraz dwie inne funkcje przedziałów ufności zwracają przedział ufności. Tak nie jest. Zwracana wartość jest połową szerokości przedziału ufności. W celu poznania pełnego przedziału ufności musisz odjąć wynik funkcji od średniej i dodać wynik do średniej.

Na rysunku 7.8 w zakresie E7:I11 został zbudowany przedział ufności identyczny z przedziałem obliczonym w zakresie E1:I4. Jest to przydatne, ponieważ pokazuje, co się dzieje za kulisami funkcji `UFNOŚĆ.NORM()`. Potrzebne są następujące obliczenia:

- Komórka F8 zawiera formułę `=F2/2`. Część pod krzywą, która jest reprezentowana przez α — tutaj 0,05, czyli 5% — musi być podzielona na pół pomiędzy dwa ogony rozkładu. Położone najbardziej po lewej 2,5% obszaru będzie umieszczone w lewym ogonie, po lewej stronie **lewego** końca przedziału ufności.
- Komórka F9 zawiera pozostałe pole powierzchni pod krzywą po usunięciu połowy α . Jest ono położone najbardziej po lewej 97,5-procentowego pola powierzchni, które się znajduje po lewej stronie **prawego** końca przedziału ufności.
- Komórka G8 zawiera formułę `=ROZKŁ.NORMALNY.S.ODWR(F8)`. Zwraca ona wartość standaryzowaną, która odcina (tutaj) położone najbardziej z lewej 2,5% pola powierzchni pod standaryzowaną krzywą normalną.
- Komórka G9 zawiera formułę `=ROZKŁ.NORMALNY.S.ODWR(F9)`. Zwraca ona wartość standaryzowaną, która odcina (tutaj) położone najbardziej po lewej 97,5% obszaru pod standaryzowaną krzywą normalną.

Teraz w komórkach G8 i G9 mamy wartości standaryzowane — liczbę odchyłeń standardowych w standaryzowanym rozkładzie normalnym — oddzielające po 2,5% rozkładu z lewej i prawej strony. W celu otrzymania tych not standardowych w jednostkach miary, których używamy — miary ilości HDL w krwi — jest niezbędne pomnożenie wartości z przez błąd standardowy średniej oraz dodanie i odjęcie tego wyniku od średniej arytmetycznej próby. Ta formuła w komórce G11 służy do obliczenia części obejmującej dodawanie:

$$=B2+(G8*C2/PIERWIASTEK(ILE.LICZB(A2:A17)))$$

W kolejności od środka na zewnątrz ta formuła przeprowadza następujące obliczenia:

1. Dzieli odchylenie standardowe w komórce C2 przez pierwiastek kwadratowy liczby obserwacji w próbie. Jak wspomniano wcześniej, ta operacja matematyczna zwraca błąd standardowy średniej.
2. Mnoży błąd standardowy średniej przez liczbę błędów standardowych poniżej średniej (–1,96), które dają w wyniku dolne 2,5% pola powierzchni pod krzywą. Ta wartość znajduje się w komórce G8.
3. Dodaje średnią próby znajdującą się w komórce B2.

Kroki od 1. do 3. zwracają wartość 46,41. Zauważ, że jest ona identyczna z dolnym limitem zwracanym przez funkcję `UFNOŚĆ.NORM()` w komórce G4.

Podobne kroki zostały zastosowane, aby otrzymać wartość w komórce I11. Różnica polega na tym, że zamiast dodawania liczby ujemnej (o wartości ujemnej wynikającej z ujemnej noty standardowej –1,96) formuła dodaje liczbę dodatnią (wartość standaryzowana 1,96 pomnożona przez błąd standardowy daje wynik dodatni). Zauważ, że wartość w komórce I11 jest identyczna z tą w I4, która zależy od funkcji `UFNOŚĆ.NORM()` zamiast od `ROZKŁ.NORMALNY.S.ODWR()`.

Zauważ, że funkcja `UFNOŚĆ.NORM()` żąda podania następujących trzech argumentów:

- **Alfa, czyli 1 minus poziom ufności** — Excel nie może przewidzieć, z jakim poziomem ufności chcesz użyć tego przedziału, więc musisz go podać.
- **Odchylenie standardowe** — ponieważ funkcja `UFNOŚĆ.NORM()` stosuje rozkład normalny do otrzymania wartości standaryzowanych związanych z różnymi polami powierzchni, zakłada, że jest to odchylenie standardowe populacji. (Więcej informacji na ten temat znajdziesz w rozdziałach 8. i 9.). Excel nie ma dostępu do pełnej populacji i dlatego nie może obliczyć jej odchylenia standardowego, a zatem ta liczba musi być podana przez użytkownika.
- **Rozmiar lub, dokładniej, liczebność próby** — nie wskazujesz samej próby (komórek A2:A17 na rysunku 7.8), więc Excel nie może zliczyć obserwacji. Musisz podać tę liczbę, aby Excel mógł obliczyć błąd standardowy średniej.

Powinieneś użyć funkcji `UFNOŚĆ.NORM()` lub `UFNOŚĆ()`, jeżeli posługujesz się nimi swobodnie i nie masz żadnego szczególnego powodu, aby wyliczać ufnosć za pomocą funkcji `ROZKŁ.NORMALNY.S.ODWR()` i błędu standardowego średniej. Po prostu pamiętaj, że funkcje `UFNOŚĆ.NORM()` i `UFNOŚĆ()` nie zwracają szerokości całego przedziału, a jedynie szerokość górnej połowy, która jest w rozkładzie symetrycznym identyczna jak szerokość dolnej połowy.

Korzystanie z funkcji `UFNOŚĆ.T()`

Rysunek 7.9 wprowadza dwie podstawowe zmiany do informacji na rysunku 7.8 — korzysta z odchylenia standardowego w komórce C2 i stosuje funkcję `UFNOŚĆ.T()` w komórce G2. Te dwie podstawowe zmiany powodują modyfikację rozmiaru wynikowego przedziału ufnosći.

Rysunek 7.9.

Przy jednakowych pozostałych parametrach przedział ufnosći zbudowany przy użyciu rozkładu t-Studenta jest szerszy niż skonstruowany za pomocą rozkładu normalnego

G2		fx		=UFNOŚĆ.T(F2;C2;ILE.LICZB(A2:A17))					
	A	B	C	D	E	F	G	H	I
		Średnia HDL	Odchylenie standardowe próby			Alfa	Połowa szerokości przedziału		
1	HDL								
2	88	57,19	20,97			0,05	11,17		
3	64								
4	50				Przedział ufności:		46,01	do	68,36
5	67								
6	45								
7	86								
8	71								
9	68								
10	36								
11	20								
12	57								
13	49								
14	37								
15	94								
16	39								
17	44								

Zauważ, że 95-procentowy przedział ufnosći na rysunku 7.9 rozciąga się od 46,01 do 68,36, natomiast na rysunku 7.8 od 46,41 do 67,97. Przedział ufnosći na rysunku 7.8 jest węższy. Uzasadnienie możesz znaleźć na rysunku 7.3. Tam możesz zobaczyć, że więcej pola powierzchni znajduje się pod ogonami rozkładu leptokurtycznego niż pod ogonami rozkładu normalnego. Musisz odejść dalej od średniej rozkładu leptokurtycznego w celu przechwycenia, powiedzmy, 95% pola powierzchni pomiędzy jego ogonami. Dlatego granice przedziału są dalsze od średniej, a przedział ufnosći jest szerszy.

Ponieważ używasz rozkładu t-Studenta, gdy nie znasz odchylenia standardowego populacji, wykorzystanie funkcji `UFNOŚĆ.T()` zamiast `UFNOŚĆ.NORM()` prowadzi do szerszego przedziału ufnosći.

Przejście od rozkładu normalnego do rozkładu t-Studenta pojawia się także w formułach w komórkach G8 i G9 na rysunku 7.9, które zawierają:

=ROZKŁ.T.ODWR(F8;ILE.LICZB(A2:A17)-1)

i

=ROZKŁ.T.ODWR(F9;ILE.LICZB(A2:A17)-1)

Zauważ, że te komórki stosują ROZKŁ.T.ODWR() zamiast ROZKŁ.NORMALNY.S.ODWR(), jak zostało to zrobione na rysunku 7.8. W dodatku do prawdopodobieństw w komórkach F8 i F9 jako argument funkcji ROZKŁ.T.ODWR() trzeba podać liczbę stopni swobody związanych z odchyleniem standardowym próby. Przypomnij sobie z rozdziału 3., że odchylenie standardowe próby stosuje w swoim mianowniku liczbę obserwacji minus 1. Gdy podasz właściwą liczbę stopni swobody, pozwolisz Excelowi użyć właściwego rozkładu t-Studenta — istnieją różne rozkłady t-Studenta dla różnych liczb stopni swobody.

Zastosowanie dodatku Analiza danych do przedziałów ufności

Dodatek *Analiza danych* Excela ma narzędzie *Statystyka opisowa*, które może być pomocne, gdy masz jedną lub wiele zmiennych do analizy. Narzędzie *Statystyka opisowa* zwraca wartościowe informacje na temat zakresu danych, w tym miary tendencji centralnej i rozproszenia oraz skośność i kurtozę. To narzędzie także zwraca połowę rozmiaru przedziału ufności, podobnie jak funkcja UFNOŚĆ.T().

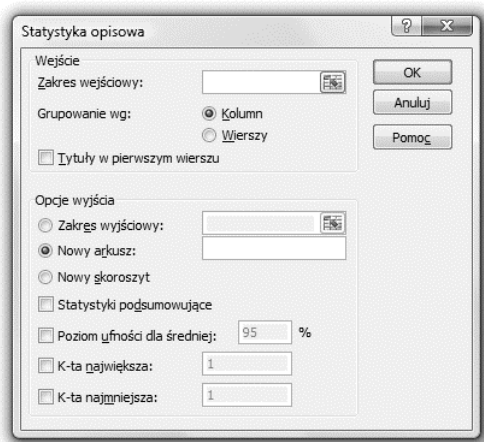
UWAGA

Obliczenia przedziału ufności za pomocą narzędzia *Statystyka opisowa* są oparte na rozkładzie t-Studenta, co jest bardzo rozsądne. Musisz podać zakres rzeczywistych danych, aby obliczyć inne statystyki opisowe, więc Excel może łatwo wyznaczyć liczebność próby i odchylenie standardowe do użycia przy wyznaczaniu błędu standardowego średniej. Ponieważ Excel oblicza odchylenie standardowe na podstawie podanego zakresu wartości, założenie, że dane stanowią próbę, jest właściwe, więc przedział ufności jest oparty na rozkładzie t-Studenta zamiast na notach standardowych.

W celu użycia narzędzia *Statystyka opisowa* musisz najpierw zainstalować dodatek *Analiza danych*. Rozdział 4. zawiera instrukcje objaśniające krok po kroku, jak zainstalować ten dodatek. Po jego zainstalowaniu z dysku pakietu Office i udostępnieniu go Excelowi znajdziesz go w grupie *Analiza* na karcie *Dane* Wstążki.

Gdy dodatek jest zainstalowany i dostępny, kliknij przycisk *Analiza danych* w grupie *Analiza* karty *Dane* i wybierz *Statystyka opisowa* z listy *Analiza danych*. Kliknij OK, aby wyświetlić okno dialogowe *Statystyka opisowa*, które zostało pokazane na rysunku 7.10.

Rysunek 7.10.
Narzędzie Statystyka opisowa jest wygodnym sposobem na szybkie otrzymanie informacji na temat miar centralnej tendencji i rozproszenia jednej lub wielu zmiennych



UWAGA

W celu obsłużenia kilku zmiennych naraz uporządkuj je w strukturę listy lub tabeli, wprowadź adres całego zakresu w pole *Zakres wejściowy* i kliknij opcję *Grupowanie wg: Kolumn*.

W celu uzyskania statystyk opisowych, takich jak średnia, skośność, liczebność itp., upewnij się, że zaznaczyłeś pole wyboru *Statystyki podsumowujące*. Aby otrzymać przedział ufności, zaznacz pole wyboru *Poziom ufności dla średniej* i wpisz w odpowiednie pole tekstowe poziom ufności, na przykład 90, 95 lub 99.

Jeżeli dane mają komórkę z nagłówkiem i uwzględniłeś ją w polu *Zakres wejściowy*, zaznacz pole wyboru *Tytuły w pierwszym wierszu*. To spowoduje, że Excel skorzysta z tej wartości jako etykiety w wynikach i nie będzie próbował używać jej jako wartości wejściowej.

Po kliknięciu przycisku *OK* otrzymasz wyniki przypominające raport pokazany na rysunku 7.11.

Rysunek 7.11.
Wyniki składają się jedynie ze stałych wartości. Nie ma tu żadnych formuł, więc po zmianie danych wejściowych nic nie zostanie obliczone automatycznie

	A	B	C	D
1	HDL		HDL	
2	88			
3	64	Średnia		57,1875
4	50	Błąd standardowy		5,242629
5	67	Mediana		53,5
6	45	Tryb		#N/D!
7	86	Odchylenie standardowe		20,97052
8	71	Wariancja próbki		439,7625
9	68	Kurtoza		-0,64987
10	36	Skośność		0,231449
11	20	Zakres		74
12	57	Minimum		20
13	49	Maksimum		94
14	37	Suma		915
15	94	Licznik		16
16	39	Poziom ufności(95,0%)		11,17
17	44			

Zauważ, że wartość w komórce D16 jest taka sama jak wartość w komórce G2 na rysunku 7.9. Wartość 11,17 jest tym, co dodajesz do średniej próby i odejmujesz od niej, aby otrzymać pełen przedział ufności.

Etykieta wyników dla przedziału ufności jest nieco myląca. Stosując standardową terminologię, **poziom ufności** nie jest wartością używaną do otrzymania pełnego przedziału ufności (tutaj 11,17). Jest to raczej prawdopodobieństwo (lub pole powierzchni pod krzywą), które wybierasz jako miarę dokładności estymacji, i prawdopodobieństwo, że przedział ufności obejmuje średnią populacji. Na rysunku 7.11 poziom ufności wynosi 95%.

Przedziały ufności i testowanie hipotez

Zarówno koncepcyjnie, jak i matematycznie przedziały ufności są blisko związane z testowaniem hipotez. Jak zobaczysz w następnych dwóch rozdziałach, często testuje się hipotezę na temat średniej próby i pewnej teoretycznej liczby lub na temat różnicy pomiędzy średnimi dwóch różnych prób losowych. W takich przypadkach mógłbyś użyć rozkładu normalnego lub blisko związanego z nim rozkładu t-Studenta, aby stwierdzić: „Hipoteza zerowa została odrzucona. Prawdopodobieństwo, że te dwie średnie pochodzą z tego samego rozkładu, jest mniejsze niż 0,05”.

To stwierdzenie jest w efekcie takie samo jak powiedzenie: „Średnia drugiej próby jest poza 95-procentowym przedziałem ufności skonstruowanym wokół średniej pierwszej próby”.

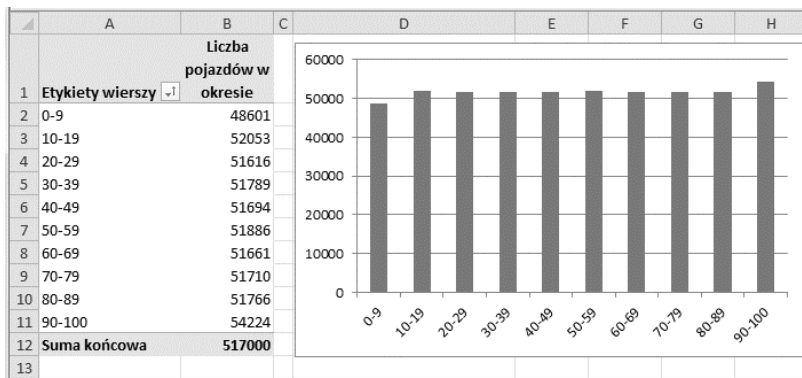
Centralne twierdzenie graniczne

Istnieje zagadnienie łączące średnią i rozkład normalny, o którym dotychczas nie wspominaliśmy. To centralne twierdzenie graniczne — groźnie brzmiące pojęcie o prostym zastosowaniu. Wygląda to tak jak w poniższej baśni.

Załóżmy, że interesuje Cię zbadanie geograficznego rozkładu ruchu pojazdów w dużym obszarze metropolitalnym. Masz nieograniczone zasoby (to właśnie aspekt baśniowy) i dlatego wysyłasz całą armię zbieraczy danych. Każdy z Twoich 2500 zbieraczy danych ma obserwować różne skrzyżowania w mieście przez sekwencje dwuminutowych okresów przez cały dzień oraz zliczać i zapisywać liczbę pojazdów, które przejeżdżają przez to skrzyżowanie w tym okresie.

Zbieracze danych wracają z całkowitą liczbą 517 000 dwuminutowych liczników pojazdów. Liczniki są dokładnie umieszczone w tabelach (kolejny aspekt baśniowy, ale to już na prawdę ostatni) i wprowadzone do arkusza Excela. Tworzysz wykresy przestawne Excela, jak pokazano na rysunku 7.12, aby otrzymać wstępny pogląd na zakres obserwacji.

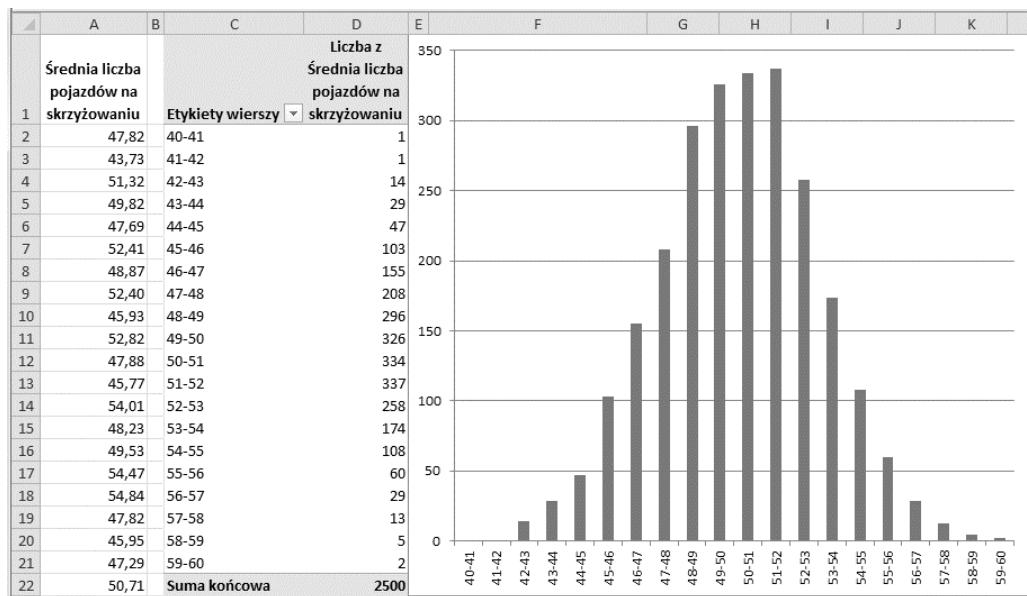
Rysunek 7.12.
Liczby samochodów
są pogrupowane
po dziesięć, aby było
łatwiej nimi zarządzać



Na rysunku 7.12 różne zakresy pojazdów są pokazane jako „etykiety wierszy” w komórkach A2:A11. Dlatego na przykład było tam 48 601 wystąpień z pomiędzy 0 a 9 pojazdami przekraczającymi skrzyżowanie w ciągu dwuminutowych okresów. Twój zbieracz danych zarejestrował inne 52 053 wystąpienia pomiędzy 10 a 19 pojazdami przekraczającymi skrzyżowanie w ciągu dwuminutowych okresów.

Zauważ, że te dane stanowią ujednolicony, regularny rozkład. Każde grupowanie (np. od 0 do 9, od 10 do 19 itd.) zawiera z grubsza tę samą liczbę obserwacji.

Następnie obliczasz i wykreślasz na wykresie **średnią** obserwację każdego z 2500 skrzyżowań. Wynik został przedstawiony na rysunku 7.13.



Rysunek 7.13. Przedstawienie średnich na wykresie przekształca rozkład prostokątny na rozkład normalny

Być może spodziewałeś się wyniku pokazanego na rysunku 7.13, a może nie. Większość osób się tego nie spodziewa. Wyjściowy rozkład jest prostokątny. Istnieje tyle samo skrzyżowań w mieście, przez które przejeżdża od zera do dziesięciu pojazdów w ciągu dwóch minut, co skrzyżowań, które są przekraczane w tym samym okresie przez od 90 do 100 pojazdów.

Ale jeżeli weźmiesz próby ze zbioru 510 000 obserwacji, obliczysz średnią każdej próby i wykreślisz te wyniki, otrzymasz coś zbliżonego do rozkładu normalnego.

I to właśnie jest nazywane **centralnym twierdzeniem granicznym**. Weź próby z populacji, której rozkład jest dowolny: prostokątny, skośny, dwumianowy, dwumodalny, jakikolwiek (na rysunku 7.12 jest on prostokątny). Oblicz średnie z każdej próby i wykreśl rozkład liczebności tych średnich (patrz rysunek 7.13). Wykres średnich będzie przypominał rozkład normalny.

Im większa liczebność próby, tym lepsza aproksymacja rozkładu normalnego. Średnie na rysunku 7.13 są oparte na próbach o liczebności 100 każda. Jeżeli próby zawierałyby, powiedzmy, po 200 obserwacji każda, wykres stałby się jeszcze bliższy rozkładowi normalnemu.

Upraszczanie spraw

Podczas pierwszej połowy dwudziestego wieku wiązano duże nadzieje z centralnym twierdzeniem granicznym jako sposobem obliczania prawdopodobieństw. Załóżmy, że badasz rozpowszechnienie leworęczności wśród graczy w golfa. Wierzysz, że 10% ogólnej populacji jest leworęczna. Dobrałeś próbę 1500 graczy w golfa i chcesz się upewnić, że nie ma żadnego systematycznego obciążenia w tej próbie. Zliczyłeś leworęcznych i znalazłeś 135. Zakładając, że 10% populacji to osoby leworęczne i że masz reprezentatywną próbę, jakie jest prawdopodobieństwo wybrania 135 lub mniej leworęcznych graczy w golfa z próby 1500?

Wzór na **dokładne** prawdopodobieństwo to:

$$\sum_{i=1}^{135} \binom{1500}{i} \cdot 0,1^i \cdot 0,9^{1500-i}$$

— możesz go zapisać, używając funkcji Excela:

```
=SUMA(KOMBINACJE(1500;WIERZSZ(A1:A135))*(0,1^WIERZSZ(A1:A135))
*(0,9^(1500-WIERZSZ(A1:A135))))
```

(Ta formuła musi być wprowadzona do Excela tablicowo, za pomocą kombinacji klawiszy **Ctrl+Shift+Enter**).

To przytłaczające, bez względu na to, czy używasz notacji sumacyjnej, czy notacji funkcji Excela. Ręczne obliczenie wyników zajęłoby długi czas, zwłaszcza że konieczne byłoby obliczenie silni 1500.

Gdy w latach 70. i 80. ubiegłego wieku rozpowszechniły się systemy mainframe i mini-komputery, obliczenie dokładnego prawdopodobieństwa stało się możliwe, choć nadal wyłącznie wtedy, gdy pracowałeś jako programista.

Od chwili pojawienia się Excela mogłeś użyć funkcji ROZKŁAD.DWUM(), a w Excelu 2010 ROZKŁ.DWUM(). Oto przykład:

=ROZKŁ.DWUM(135;1500;0,1;PRAWDA)

Dowolna z tych formuł zwraca dokładne prawdopodobieństwo w rozkładzie dwumianowym równe 10,48%. (Ta liczba może, ale nie musi spowodować, że zdecydujesz, iż Twoja próba jest niereprezentatywna — jest to decyzja subiektywna). W 1950 roku nie było jednak jeszcze dostępnej odpowiedniej mocy obliczeniowej. Musiałeś polegać na suwakach logarytmicznych i kompilacjach tabel matematycznych oraz naukowych, aby wykonać całą pracę i skończyć z wynikiem bliskim liczby 10,48%.

Alternatywnie mogłeś przywołać centralne twierdzenie graniczne. Pierwszą sprawą jest zauważenie, że zmienna dychotomiczna, taka jak preferowana ręka — praworęczność albo leworęczność — ma odchylenie standardowe, podobnie jak dowolna zmienna liczbowa. Jeżeli uznasz, że p oznacza jedną proporcję, na przykład 0,1, a $(1 - p)$ oznacza drugą proporcję (0,9), wtedy odchylenie standardowe tej zmiennej jest następujące:

$$\sqrt{p(1 - p)}$$

Jest to pierwiastek kwadratowy iloczynu tych dwóch proporcji, których suma wynosi 1,0. Z próbą pewnej liczby N osób, które mają daną cechę lub jej nie mają, odchylenie standardowe tej liczby osób jest następujące:

$$\sqrt{Np(1 - p)}$$

a odchylenie standardowe rozkładu preferowanej ręki 1500 graczy w golfa, przy założeniu 10% leworęcznych i 90% praworęcznych, wyniosłoby:

$$\sqrt{1500 \cdot 0,1 \cdot 0,9}$$

czyli 11,6.

Wiesz, że w próbie losowej liczba graczy w golfa, którzy są leworęczni, powinna wynosić 10% z 1500, czyli 150. Znasz odchylenie standardowe równe 11,6. A centralne twierdzenie graniczne określa, że średnie wielu prób losowych są zgodne z rozkładem normalnym przy założeniu, że te próby są wystarczająco duże. 1500 jest z pewnością dużą próbą.

Dlatego powinienes móc porównać swój wynik 135 leworęcznych graczy w golfa z rozkładem normalnym. Obserwowana liczba 135, pomniejszona o średnią 150 i podzielona przez standardowe odchylenie równe 11,6, zwraca wartość standaryzowaną równą -1,29. W dowolnej tabeli prezentującej pola powierzchni pod krzywą normalną — a te są w dowolnych podstawowych podręcznikach statystyki — znajdziesz informację, że wartość z równa -1,29 odpowiada polu powierzchni, czyli prawdopodobieństwu 9,84%. W przypadku braku podręczników statystycznych możesz użyć formuły:

=ROZKŁ.NORMALNY.S(-1,29;PRAWDA)

albo równoważnie:

=ROZKŁ.NORMALNY(135;150;11,6;PRAWDA)

Wynikiem zastosowania rozkładu normalnego jest 9,84%. Wynik dokładnego rozkładu dwumianowego to 10,48 — różnica wynosi odrobinę powyżej połowy punktu procentowego.

Ulepszanie spraw

Liczba 9,84% jest nazywana „normalnym przybliżeniem rozkładu dwumianowego”. Była i w pewnym stopniu pozostała popularną alternatywą dla użycia samego rozkładu dwumianowego. Swą popularność zawdzięcza temu, że obliczenie formuły kombinacji C_n^k było pracochłonne i podatne na błędy. To przybliżenie jest nadal czasem używane, ponieważ nie każdy, kto od połowy lat 80. ubiegłego wieku chciał obliczyć prawdopodobieństwo dwumianowe, miał dostęp do odpowiedniego oprogramowania. Stąd pewna poznawcza bezsilność, z którą się nadal borykamy.

Drobną rozbieżność pomiędzy 9,84% a 10,48% statystycy uznawali przed laty za nieistotną — i prawdopodobnie taka jest. Jednak w przypadku przybliżania normalnego pojawiają się inne ograniczenia — odradza się na przykład używania go, gdy wartość Np albo $N(1-p)$ jest mniejsza niż 5 czy, jak podają inne źródła, niż 10. Trwają spory na temat użycia „poprawki na ciągłość”, która pozwala analizować takie wartości jak liczba graczy w golfa rosnąca co 1 (nie możesz mieć 3/4 gracza w golfa). Dlatego normalne przybliżenie rozkładu dwumianowego, przed dostępnością dużej mocy obliczeniowej, którą się teraz cieszymy, było błogosławieństwem przyjmowanym z dość mieszanymi odczuciami.

Normalna aproksymacja rozkładu dwumianowego jest możliwa dzięki zastosowaniu centralnego twierdzenia granicznego. Ponieważ jednak prawdopodobieństwo dwumianowe stało się względnie łatwe do obliczenia, normalne aproksymacje rozkładów dwumianowych są coraz rzadziej spotykane. To samo dotyczy innych przybliżeń. Centralne twierdzenie graniczne pozostaje kamieniem węgielnym teorii statystycznej, ale (już w 1970 r.) powszechnie poważany statystyk napisał: „że nie odgrywa ono takiej istotnej roli jak kiedyś”.

Skorowidz

α , 254

β , 254

A

akceptacja hipotezy, 149

alfa, 150, 184

algebra macierzowa, 129

analiza całkowitego rozproszenia, 294

analiza czynnikowa, 321

analiza kowariancji (ANCOVA), 279, 403, 408, 424, 429

analiza kowariancji wielorakiej, 441

analiza regresji, 113, 174, 180

analiza regresji wielorakiej, 224

analiza rozkładu dwumianowego, 147

analiza statystyczna, 180

analiza testów psychologicznych, 180

analiza wariancji (ANOVA), 174, 224, 273, 293, 343, 352

podstawy analizy, 294

porównywanie wariancji, 297

Test F, 183, 301, 311

analiza wariancji dla regresji, 363

analiza wariancji i kowariancji, 183

analiza z powtarzanymi obserwacjami, 338

ANCOVA, 279, 403, 408, 424, 429

ANOVA, 174, 224, 273, 293, 343, 352

argument funkcji, 52

argument ślady, 270, 272

argument typ

grupy zależne, 282

nierówne wariancje, 282

równe wariancje, 282

argument typ funkcji T.TEST(), 282

argumenty funkcji ROZKŁ.F(), 305

arkusz ukryty, 46

asymetria rozkładu skośnego, 194

automatyczne rozszerzanie formuły, 374

B

badanie interakcji, 323

badanie linii regresji, 421

badanie obserwacyjne, 389

blok zrandomizowany, 338

błąd, 31

błąd $\#N/D!$, 65, 391

błąd e, 355

błąd próbkowania, 185, 257

błąd resztowy, 404, 434

błąd standardowy, 227

błąd standardowy estymacji, 394

błąd standardowy różnicy, 260

błąd standardowy różnicy pomiędzy średnimi, 281

błąd standardowy średniej, 226, 227, 245, 280

błąd standardowy średniej arytmetycznej, 209

błąd typu I, 230

błąd wariancji średniej, 227, 300, 352

budowanie rozkładu liczebności, 37

C

Campbell Donald, 175

cel korygowania średnich, 425

cele analizy kowariancji, 404

centralne twierdzenie graniczne, 217

czynnik, 296, 321

czynnik losowy, 342

czynnik różnicujący, 329

czynnik stały, 342

czynniki główne, 329

czynnikowa analiza wariancji, 321

D

definicja kurtozy, 199

definicja wariancji, 85

definicja współczynnika korelacji, 103

definiowanie reguły decyzyjnej, 242

diagram korelacyjny, 29

dodanie zmiennej towarzyszącej, 442

dodatek

Analiza danych, 107, 174, 215, 282

Analysis ToolPak, 107, 173

Solver, 57

dodatkowy poziom czynnika, 419

dodawanie zmiennej towarzyszącej, 408

dokładne prawdopodobieństwo, 219

dokładność prognozowania, 123

dokładność prognozowanych wartości, 124

dokładność regresji, 415

dokumentacja dodatku, 174, 181

dominanta, 50

druga zmienna towarzysząca, 442

E

efekt Yule'a i Simpsona, 162

eksperyment, 389

eksperyment prawdziwy, 398

eliminacja wsteczna, 389

estymator, 91

estymator MS_b, 301estymator MS_w, 301

estymator nieobciążony, 92

estymator standardowego odchylenia, 210

estymator wariancji populacji, 301

F

faza projektowania eksperymentu, 175

faza wdrażania eksperymentu, 175

format walutowy, 56

formuła, 52, 54

formuła tablicowa, 39

Ctrl+Shift+Enter, 40

nawiasy klamrowe, 70

zastosowania, 71

formuła tablicowa do zliczania wartości, 70

funkcja

CHI.TEST(), 156, 160, 169

CZĘSTOŚĆ(), 38

gęstości prawdopodobieństwa, 202

gęstości rozkładu normalnego, 35

ILE.LICZB(), 258

JEŻELI(), 70

KURTOZA(), 199

LICZ.JEŻELI(), 75

LOS(), 143

MACIERZ.ILOCZYN(), 130, 316

MACIERZ.ODW(), 130, 349

MEDIANA(), 62

MODUŁ.LICZBY(), 320

NACHYLENIE(), 120

ODCH.KWADRATOWE(), 258, 296, 331

ODCH.STAND.POPUL(), 91

ODCH.STANDARD.POPUL(), 91

ODCH.STANDARD.PRÓBK(), 91

ODCH.STANDARDOWE(), 85, 91

ODCIĘTA(), 120

PEARSON(), 98

PODAJ.POZYCJĘ(), 69

R.KWADRAT(), 370, 373

REGLINP(), 120, 123, 125, 368, 390, 429

błąd standardowy, 392

błąd standardowy estymacji, 394

wartość F, 395

wieloraki R², 394

wiersze od trzeciego do piątego, 394

współczynniki regresji, 391

wyraz wolny, 392

REGLINW(), 117, 365, 367, 368, 373

jako formuła tablicowa, 119

nowe_x, 118

zmienne objaśniające, 124

znane_x, 118

znane_y, 118

ROZKŁ.CHI(), 166

ROZKŁ.CHI.ODWR(), 167

prawdopodobieństwo, 167

stopnie_swobody, 167

ROZKŁ.CHI.ODWR.PS(), 168

ROZKŁ.CHI.PS(), 162, 166

ROZKŁ.DWUM(), 137

interpretacja wyników, 139

liczba sukcesów, 138

prawdopodobieństwo sukcesu, 139

próby, 138

skumulowany, 139

ROZKŁ.DWUM.ODWR(), 145

liczba sukcesów, 145

prawdopodobieństwo sukcesu, 145

próby, 145

skumulowany, 145

ROZKŁ.F(), 188, 305
 ROZKŁ.F.ODWR(), 305, 306
 ROZKŁ.F.ODWR.PS(), 189, 317, 439
 ROZKŁ.F.PS(), 188, 305
 ROZKŁ.NORMALNY(), 200, 231
 odchylenie standardowe, 200
 skumulowany, 201
 sposoby użycia funkcji, 201
 średnia, 200
 x, 200
 ROZKŁ.NORMALNY.ODWR(), 203, 204, 244
 ROZKŁ.NORMALNY.S(), 204, 235
 ROZKŁ.NORMALNY.S.ODWR(), 205
 ROZKŁ.T(), 264
 skumulowany, 264
 stopnie_swobody, 264
 x, 264
 ROZKŁ.T.DS(), 264
 ROZKŁ.T.ODWR(), 245, 261
 ROZKŁ.T.PS(), 264
 ROZKŁAD.CHI(), 166
 stopnie_swobody, 166
 x, 166
 ROZKŁAD.DWUM(), 138, 220
 ROZKŁAD.F(), 306
 ROZKŁAD.F.ODW(), 306
 ROZKŁAD.NORMALNY.ODW(), 204
 ROZKŁAD.T.ODW(), 261
 SKOŚNOŚĆ(), 195
 SUMA(), 71
 SUMA.ILOCZYNÓW(), 316
 SUMA.JEŻELI(), 75
 ŚREDNIA(), 51, 61
 T.TEST(), 265, 269, 282
 interpretacja wyniku, 271
 ślady, 270
 tablice, 269
 typ, 275
 TEST.CHI(), 169
 UFNOŚĆ(), 211
 UFNOŚĆ.NORM(), 211, 212
 alfa, 213
 odchylenie standardowe, 213
 rozmiar, 213
 UFNOŚĆ.T(), 211, 212, 214

WARIANCJA(), 85
 WARIANCJA.POPUL(), 91
 WSP.KORELACJI(), 97, 103
 WYST.NAJCZĘŚCIEJ(), 63, 66
 WYSZUKAJ.PIONOWO(), 357, 358
 funkcje, 52
 argument, 52
 zwracanie wyniku, 54
 funkcje Excela, 339
 funkcje macierzowe, 130
 funkcje odchylenia standardowego, 93
 funkcje wariancji, 94
 funkcje χ^2 , 164

G

główna przekątna macierzy korelacji, 379
 główny czynnik, 380
 grupa, 39
 grupa eksperymentalna, 176
 grupa kontrolna, 176
 grupowanie
 funkcja CZĘSTOŚĆ(), 38
 tabela przestawna, 42

H

hipoteza, 136
 hipoteza alternatywna, 136, 225
 hipoteza dwustronna, 190, 254
 hipoteza jednostronna, 190, 254, 256
 hipoteza kierunkowa, 190, 254
 hipoteza zerowa, 136, 157, 224, 301
 Huff Darrell, 173

I

iloraz dwóch wariancji, 302
 iloraz t, 320, 441
 indeks tabeli przestawnej, 171
 instalowanie narzędzia Solver, 57
 interakcja, 328
 interakcja czynnika i zmiennej towarzyszącej, 420
 istotność różnic średnich w grupach, 444
 istotność statystyczna, 329

J

jednorodność współczynników regresji, 416

K

kategoria, 24
klasa, 39
klasowy rozkład liczebności, 38
klawisz ponownego wyliczania F9, 75
kodowanie ortogonalne, 349
kodowanie z użyciem zmiennych sztucznych, 348
kody grup, 347
kolejność wpisów, 381, 384
komórka, 325
komórka arkusza, 325
komórka schematu, 326
konstruowanie równania regresji, 390
korekta średniego kwadratu reszty, 436
korekty wartości średnich, 432
korelacja, 95, 111, 280
 problem złożoności związków, 112
 trzecia zmienna, 112
 zastosowania, 113
korelacja dodatnia, 96
korelacja mocna, 106
korelacja semicząstkowa, 364, 365
korelacja silna, 112
korelacja słaba, 106
korelacja ujemna, 96
korygowanie średnich, 425
koszt doświadczenia, 47
koszt testów dwustronnych, 273
kowariancja s_{xy} , 99, 115, 127
kowariancja bezwymiarowa, 115
krzywa dzwonowa, 194
krzywa Gaussa, 194
krzywa normalna, 35
krzyżowanie, 321
kształt niecentralnego rozkładu F, 341
kurtoza, 196
kwadrat współczynników korelacji, 349, 385
kwadraty odchyleń, 61

L

lewy ogon rozkładu, 255
lewy ukośnik, 75
liczba resztowych stopni swobody, 440
liczba stopni swobody, 89–93, 100, 162, 240, 266, 334, 395, 397, 418
liczba stopni swobody regresji, 437
liczba stopni swobody wariacji wewnątrzgrupowej, 297
liczba wektorów, 347
liczby pseudolosowe, 143
liczebność grup, 266, 309, 335
liczebność klasy, 38
liczebność próby losowej, 36, 194, 209
liczebność próby N, 259
licznik ilorazu t, 440
linia regresji, 100, 114, 411
linia trendu, 30, 100, 411
lista, 20
losowy wybór próby, 141, 142

Ł

łączenie predyktorów, 123

M

macierz korelacji, 110, 379
maksymalizacja R2, 389
mała liczebność próby, 241
margines błędu, 36
mediana, 49, 61
metoda najmniejszych kwadratów (MNK), 31
metoda Newmana-Keulsa, 312
metoda planowanych różnic ortogonalnych, 320
metoda Scheffégo, 434
miara nieliniowej korelacji, 105
miara rozproszenia
 odchylenie standardowe, 82
 roztęp, 78
 wariancja, 85
miara skośności, 196
miara tendencji centralnej, 63
 mediana, 61
 moda, 63
 średnia arytmetyczna, 51

minimalizacja sumy, 61
minimalizowanie kwadratów odchyłeń, 61
minimalizowanie rozproszenia, 56
mit gracza, 154
moc statystyczna testu, 246, 323, 405, 440
moc statystyczna testu t, 247, 289
moc statystyczna testu F, 341, 445
moda, 63
model Bayesa, 141
model mieszany, 342
model ograniczony, 427

N

N–1, liczba stopni swobody, 89–93, 100, 162, 240, 266, 334, 395, 397, 418
nachylenie (współczynnik kierunkowy) linii regresji, 120, 422, 428
nagłówków, 21
najmniejsze kwadraty, 56
narzędzia Excela, 339
narzędzie
 Analiza danych, 111
 Analiza wariancji, 302, 313, 323–326, 338, 407, 357
 F-Test z dwiema próbami dla wariancji, 174
 Korelacja, 107, 108, 109, 174
 macierz współczynników korelacji, 110
 zakres wejściowy, 109
 zakres wyjściowy, 109
 Kowariancja, 174
 Regresja, 126, 344, 367
 Solver, 57
 Statystyka opisowa, 215
 Szacowanie formuły, 73
 Szukanie wyniku, 57
 Test F, 181, 185, 268
 hipoteza dwustronna, 191, 192
 hipoteza kierunkowa, 192
 interpretowanie wyniku, 185
 wariancje dwóch prób losowych, 182
 wartość p, 191
 wykres, 190
 Test F z dwiema próbami dla wariancji, 183
 Test t, 267
nawiasy klamrowe, 40

nazywanie zakresów, 257
niecentralny rozkład F, 341
niedoszacowanie błędu standardowego, 268
niedoszacowanie wariancji, 89
niejednoznaczność, 389
nieobciążenie, 91
nierówne liczebności, 337
nierówne liczebności grup, 309, 360, 379, 397, 399
nierówne N, 337
nierówne wariancje, 267
niezależność klasyfikacji, 156
niezależność wyborów, 143
N–J, 297
notacja macierzowa, 130

O

obciążenie estymatora, 92
obciążenie wyboru, 177
obliczanie
 błędu standardowego grup zależnych, 277
 błędu standardowego różnic średnich, 259
 efektu interakcji, 330
 kurtozy, 198
 mediany, 61
 oczekiwanych częstości, 170
 odchylenia standardowego, 85
 planowanych różnic ortogonalnych, 319
 prawdopodobieństwa, 281
 przedziału ufności, 215
 rozstępu, 80
 skorygowanych średnich, 412
 skośności, 196
 statystyki t, 260, 281
 statystyki χ^2 , 161
 sumy kwadratów, 350
 sumy kwadratów odchyłeń, 257
 średniej arytmetycznej, 51
 wariancji, 85
 wariancji sumarycznej, 258
 wartości krytycznej, 261
 wartości modalnej, 63
 wartości α , 304
 współczynnika korelacji, 97
odchylenia standardowe różnic, 437
odchylenie, 61

- odchylenie grupy, 300
- odchylenie standardowe, 81, 82, 88
- odchylenie standardowe populacji σ , 88
- odchylenie standardowe próby losowej s , 88
- odległość obserwacji od średniej grupy, 355
- odległość wyników od średniej, 56
- odrzućcie hipotezy, 149
- odrzućcie hipotezy zerowej, 159, 246
- odwołanie bezwzględne, 374
- odwołanie mieszane, 374
- odwołanie względne, 374
- ogólna suma kwadratów, 431
- ogólny iloczyn wektorowy, 432
- ograniczenia argumentów funkcji, 402
- określenie poziomu α , 191
- oś kategorii, 24
- oś wartości, 24

P

- paradoks Simpsona, 163
- parametr, 88
- parametr niecentralności, 342
- parametry populacji, 355
- pasek formuły, 55
- pasek stanu, 60
- Pearson Karl, 114
- pełen model, 427
- plan z powtarzanymi obserwacjami, 339
- planowana różnica, 440
- planowane różnice ortogonalne, 317
- podejście regresyjne, 410
- podział całkowitego rozproszenia, 297
- pole, 20
- pole formuły, 28
- pole nazwy, 257
- polecenie Grupuj, 43
- polecenie Liniowa linia trendu, 120
- poprawka na ciągłość, 221
- populacja, 89, 225
- populacja o rozkładzie normalnym, 46
- porównania wielokrotne, 293, 311, 434
 - planowane różnice ortogonalne, 317
- procedura Scheffégo, 311, 312, 320
- porównania wielokrotne a priori, 435

- porównania wielokrotne post hoc, 435
- porównanie
 - skorygowanych średnich, 430
 - wartości R^2 , 410
 - statystyk, 25
 - wariancji, 297
- porządkowanie a priori, 390
- powtórzenia, 326, 338
- poziom istotności, 230, 251
- poziom ufności, 36, 217, 445
- poziom α , 230, 251
- prawdopodobieństwo skumulowane, 165
- prawdopodobieństwo w rozkładzie dwumianowym, 144
- prawdziwa obserwowana średnia, 413
- prawy ogon lewego rozkładu, 261
- prawy ogon lewej krzywej, 274
- prawy ogon rozkładu, 255
- predyktor, 123, 346
- problem Behrensa-Fishera, 163, 310, 387
- problem z kolejnością wprowadzania zmiennych, 398
- procedura
 - Dunna, 311
 - Dunnetta, 311
 - porównań wielokrotnych, 311
 - Scheffégo, 311, 312, 320
 - Tukeya i Newman-Keuls, 311
- prognozowanie wartości, 118
 - dokładność, 124
 - formuła z funkcją REGLINW(), 118
 - przy użyciu równania regresji, 125
- program R, 312
- proporcjonalne liczebności komórek, 338
- próba losowa, 37, 89
- próba z populacji, 46
- przedział, 39
- przedział ufności, 205
 - dla średniej, 207
 - funkcja ufność, 211
- narzędzie Statystyka opisowa, 215
- rozkład normalny, 211
- rozkład t-Studenta, 214
- rozszerzanie przedziału, 208

R

raport Analizy wariancji, 303
 redukcja obciążenia, 405
 regresja, 113
 regresja krokowa, 389
 regresja logistyczna, 294
 regresja mniej dokładna, 414
 regresja w stronę średniej, 178
 regresja wieloraka, 123, 344, 360
 rekodowanie, 348
 rekodowanie zmiennych, 346–347, 355, 362, 431
 rekord, 20
 reszta, 421
 resztowe stopnie swobody, 430
 reszty regresji, 370
 rozkład dwumianowy, 136, 221
 wzór na prawdopodobieństwo, 144
 rozkład F, 195, 302, 308
 wykres, 308
 rozkład liczebności, 31, 87
 rozkład liczebności dodatnio skośny, 33
 rozkład liczebności przesunięty, 33
 rozkład liczebności symulowany, 45
 rozkład liczebności średnich, 219
 rozkład liczebności ujemnie skośny, 34
 rozkład normalny, 35, 82, 193
 rozkład odniesienia, 64
 rozkład populacji, 45
 rozkład próbkowania, 137
 rozkład skośny, 63
 rozkład t, 320
 rozkład t-Studenta, 194, 214, 242, 255, 263
 rozkład χ^2 , 156, 157
 rozkład χ^2 z n stopniami swobody, 158
 rozmiar marginesu błędu, 36
 rozproszenie, 78, 294
 rozproszenie grup, 280
 rozproszenie pomiędzy grupami, 300
 rozróżnienie korelacji i przyczynowości, 112
 rozrzut, 77
 rozstęp, 78, 79
 rozstęp studentyzowany, 312
 równanie regresji, 125
 różnice ortogonalne, 318

S

schemat eksperymentalny, 387
 schemat niezrównoważony, 379, 380
 schemat zrównoważony, 362, 378, 379
 schematy czynnikowe, 362
 selekcja postępująca, 389
 siła związku, 101
 skala ilorazowa (stosunkowa), 26
 skala liczbowa, 25
 ilorazowa, 25
 porządkowa, 25
 przedziałowa, 25, 402
 skala nominalna, 23
 skala przedziałowa (interwałowa), 25, 402
 skorygowane średnie, 411, 413
 skorygowane średnie grup, 423, 431
 skośność, 194
 skrót
 df, 303
 MS, 303
 SS, 303
 SV, 303
 skumulowane prawdopodobieństwo, 201, 292
 sposób wnioskowania, 34
 sprawdzanie hipotez, 148
 standaryzowany rozkład normalny, 199
 Stanley Julian, 175
 statystyka, 88
 statystyka rozstępu studentyzowanego, 311
 statystyka t., 241
 statystyki funkcji REGLINP(), 390
 statystyki opisowe, 35, 175, 436
 stopa błędu α , 242
 stopa błędu β , 248
 stopnie swobody (df), 92, 265
 stopnie swobody pomiędzy grupami, 437
 stosowanie planowanych różnic, 439
 strata stopni swobody z reszt, 445
 suma kwadratów pomiędzy grupami, 295, 352
 suma kwadratów wewnątrz grup, 296, 297, 352
 suma kwadratów zmiennych towarzyszących, 444
 symetria połączona, 339
 szacowanie wariancji
 za pomocą analizy wariancji, 352
 za pomocą regresji, 353
 szerokość przedziału ufności, 211, 214

Ś

średni błąd kwadratowy, 315, 355
 średni kwadrat pomiędzy grupami, 298, 345
 średni kwadrat regresji, 345
 średni kwadrat resztowy, 345, 436
 średni kwadrat wewnątrz grup, 298, 345
 średni współczynnik regresji, 413–416
 średnia, 51
 średnia arytmetyczna, 49, 51
 średnia linii regresji, 416
 średnia populacji μ , 90
 średnia próba losowa, 90
 średnia zmiennej towarzyszącej, 413
 średnie grup, 280
 średnie odchylenie, 88
 średnik, 75

T

tabela, 21
 tabela analizy wariancji, 436
 tabela kontyngencji (krzyżowa), 152, 163
 tabela przestawna, 21, 42, 68
 tabela przestawna dwuwymiarowe, 151
 tabela przestawna jednowymiarowe, 133
 tablicowe wprowadzanie formuły, 39, 72
 tendencja centralna, 75
 teoretyczny rozkład χ^2 , 159
 teoria rozkładów dwumianowych, 138
 termin
 analiza kowariancji, 404
 interakcja, 328
 kowariancja wieloraka, 441
 powtórzenia, 326
 regresja wieloraka, 345
 zrandomizowany, 338
 test Boxa, 339
 test dwustronny, 273
 test F, 183, 301, 311
 test F Greenhouse'a-Geissera, 339
 test F konserwatywny, 310
 test F liberalny, 310
 test F omnibusa, 436
 test F z dwiema próbami dla wariancji, 181
 test jednorodności współczynników regresji, 421

test Newmana-Keulsa, 312
 test par skojarzonych, 285
 test średnich, 223
 test t, 240, 253, 257
 analiza formuły, 279
 założenia na temat danych źródłowych, 275
 test t grup niezależnych, 278
 test t grup zależnych, 266, 277, 283
 test t konserwatywny, 283
 test t liberalny, 283
 test t zakładający nierówne wariancje, 286
 test t zakładający równe wariancje, 283
 test z, 225
 test χ^2 , 156
 testowanie hipotez, 217
 testowanie hipotezy zerowej, 160
 testowanie różnic pomiędzy średnimi, 223, 253, 408
 analiza wariancji, 291
 dodatek Analiza danych, 282
 funkcja T.TEST(), 265
 funkcje ROZKŁ.T() i ROZKŁ.T.ODWR(), 254
 testowanie różnic średnich grup, 182
 testowanie średniego współczynnika regresji, 416
 testy, 180
 trafność, 175
 trafność wewnętrzna, 176
 zagrożenia, 177
 zapewnienie, 176
 trend, 100
 tworzenie kombinacji liniowej, 124
 tworzenie tabeli przestawnej, 134
 tworzenie wykresów, 232, 236
 odchylenia standardowe, 234
 oś pozioma, 233
 rozkład średnich prób losowych, 236
 średnia próby losowej, 236
 wartości populacji, 234
 wartości standaryzowane, 233
 tworzenie wykresu XY, 106
 typy kurtozy, 197

U

uchwyt zaznaczania, 374
 układ danych, 427
 układ listy, 426

unikalne części wariancji, 397
uporządkowanie danych, 20
ustalanie reguł decyzyjnych, 140
usuwanie
 obciążenia, 418
 wpływu predyktora, 366
 wpływu zmiennej, 365

V

VBA, Visual Basic for Applications, 46, 341, 389

W

wariancja, 83, 85, 297
wariancja błędu, 355
wariancja dwóch połączonych grup, 259
wariancja populacji σ^2 , 89, 226, 299
wariancja próby losowej s^2 , 89
wariancja resztowa, 353
wariancja sumaryczna, 257
wariancja wewnątrzgrupowa, 259
wariancje grup, 267, 283
wartość, 24
 bj, 356
 centralna, 50
 F, 186, 189, 304
 hipotetyczna, 255
 krytyczna, 244
 krytyczna F, 304
 krytyczna statystyki t, 257
 krytyczna testu t, 245
 krytyczna testu z, 244
 modalna, 50, 63, 64, 68
 modalna z formuły, 69
 prognozowana, 123
 R2, 410
 resztowa, 370
 standaryzowana, 82, 84, 224
 standaryzowana z, 114, 158
 t, 245
 tekstowa, 27
 zmiennej, 20
wektor, 346
wektor interakcji, 380
wektory kodowe, 354, 357

wiele zmiennych towarzyszących, 442
wieloraki R2, 128, 351
wizualizacja mocy statystycznej, 288
wizualizacja rozkładu, 34
wnioskowanie statystyczne, 36, 175
wpływ, 356
współczynnik determinacji, 128
współczynnik korelacji r, 97, 122, 379
 liniowość, 105
 rodzaje pomyłek, 104
współczynnik korelacji częściowej, 365
współczynnik korelacji semicząstkowej, 365
współczynnik różnicy, 314
współdzielona zmienność, 127
wstępna analiza wariancji, 313
wykres
 kolumnowy, 23
 Kolumnowy grupowany, 40
 liniowy, 27
 przestawny, 21, 39, 67
 punktowy, 29
 rozkładu liczebności, 32
 rozrzutu, 29
 słupkowy, 24
 XY (punktowy), 27, 29
wykresy przestawne, 217
wynik formuły, 54
wyraz wolny, 121, 392
wyregulowanie testu t, 277

Z

zaawansowane metody statystyczne, 294
zagnieżdżanie, 321
założenie losowości próby, 141
założenie o niezależności wyborów, 143
założenie o niezależności rekordów, 276
założenie o normalnym rozkładzie, 276
założenie o równości wariancji, 182, 267, 283
zaznaczenie zakresu danych, 109
zdarzenia niezależne, 154
zliczanie próby losowej, 37
zmienna, 20
zmienna dychotomiczna, 220
zmienna ilościowa, 25
zmienna nominalna, 25

zmienna objaśniająca, 123, 341
zmienna objaśniana, 407
zmienna towarzysząca, 403, 407
zmienna wskaźnikowa, 346
zmienne nie skorelowane, 379
zmienne nominalne, 133, 344
zmienne skorelowane, 380
zmienność, 352, 442
znak dolara, 122
zrównoważony schemat czynnikowy, 387
związek przyczynowo-skutkowy, 112
zwracanie wyniku, 54

PROGRAM PARTNERSKI

GRUPY WYDAWNICZEJ HELION



1. ZAREJESTRUJ SIĘ
2. PREZENTUJ KSIĄŻKI
3. ZBIERAJ PROWIZJĘ

Zmień swoją stronę WWW
w działający bankomat!

Dowiedz się więcej i dołącz już dzisiaj!

<http://program-partnerski.helion.pl>

GRUPA WYDAWNICZA



Helion SA

Analiza statystyczna. Microsoft® Excel 2010 PL

Microsoft Excel 2010 to ukochane narzędzie studentów, analityków, księgowych, menedżerów i prezesów. Uniwersalność i niezawodność przy wykonywaniu operacji na ogromnych zbiorach danych zapewniła mu popularność w wielu dziedzinach życia. Jedną z nich jest statystyka i wnioskowanie statystyczne.

Dzięki tej książce wykorzystasz potencjał Excela do wyciągania lepszych, dokładniejszych i bardziej wiarygodnych wniosków na podstawie dostępnych danych. Nauczysz się graficznie prezentować informacje, tworzyć zaawansowane formuły oraz korzystać z niezwykle przydatnego narzędzia Solver. Ponadto dowiesz się, jak określić korelację zmiennych, korzystać z testów oraz nowych funkcji spójności. Książka ta idealnie sprawdzi się w rękach studentów mających do czynienia ze statystyką podczas zajęć. Zachwyci także menedżerów, których decyzje mogą zaważyć na losach firm i prowadzonych przez nie projektów!

Dzięki tej książce:

- biegle opanujesz funkcje statystyczne aplikacji Microsoft Excel 2010
- będziesz wyciągał celne wnioski na podstawie posiadanych danych
- bez problemu zbadasz korelację i regresję zmiennych
- wykorzystasz narzędzie Solver oraz dodatkowe narzędzia z dodatku *Analiza danych Excela*

Trafne decyzje w zasięgu ręki!

helion.pl
księgarnia
internetowa

Nr katalogowy: 7814



Księgarnia internetowa:
<http://helion.pl>



Zamówienia telefoniczne:

0 801 339900



0 601 339900



Helion

Sprawdź najnowsze promocje:

• <http://helion.pl/promocje>

Książki najchętniej czytane:

• <http://helion.pl/bestsellery>

Zamów informacje o nowościach:

• <http://helion.pl/nawosci>

Helion SA

ul. Kościuszki 1c, 44-100 Gliwice

tel.: 32 230 98 63

e-mail: helion@helion.pl

<http://helion.pl>

sięgnij po **WIĘCEJ**



KOD KORZYŚCI

ISBN 978-83-246-3668-6



9 788324 636686

Cena: 79,00 zł

Informatyka w najlepszym wydaniu