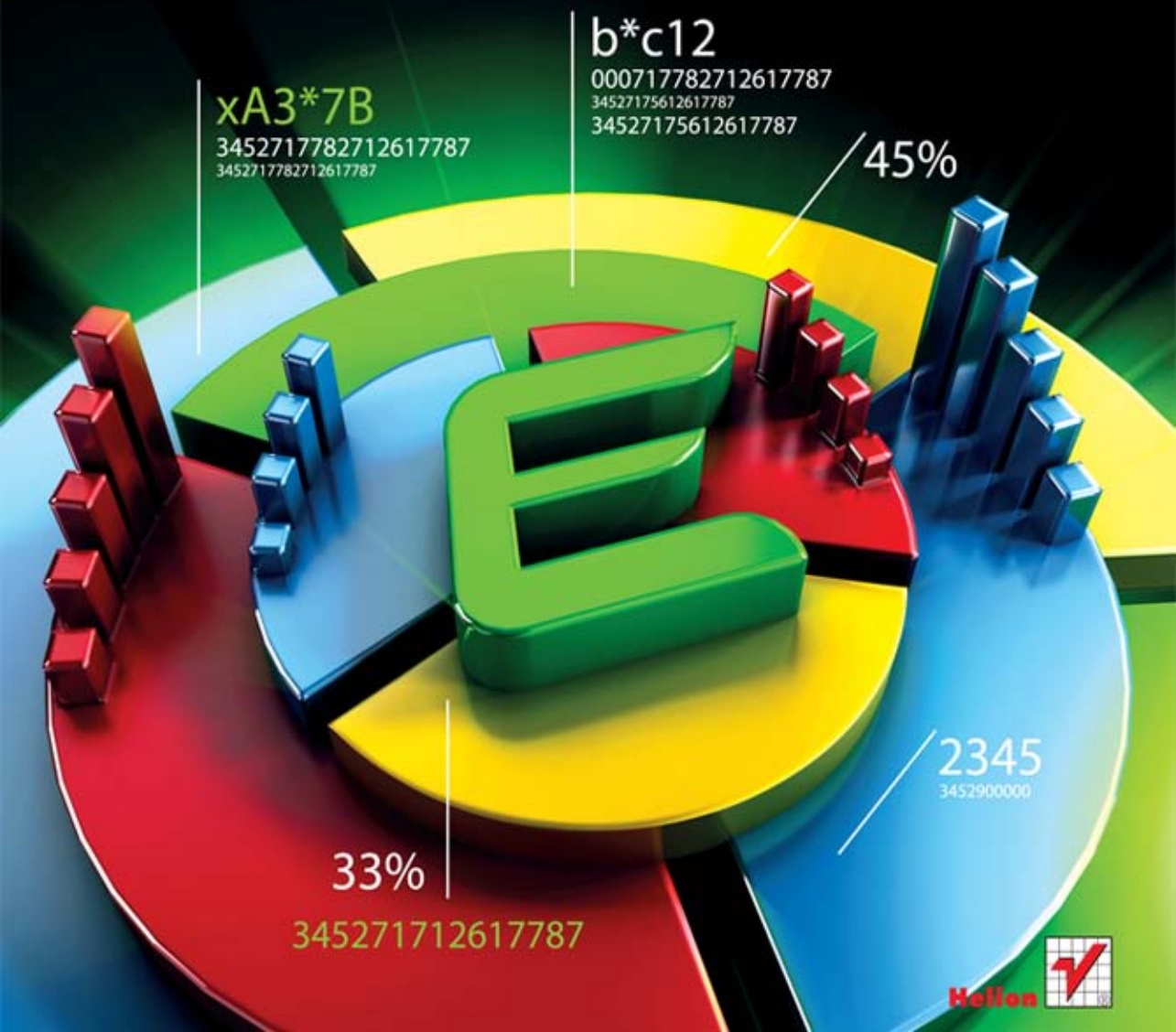
Zaufaj posiadanym danym!

Conrad Carlberg

Analiza statystyczna. Microsoft® Excel 2010 PL



Helion

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.

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.

Spis treści

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:

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

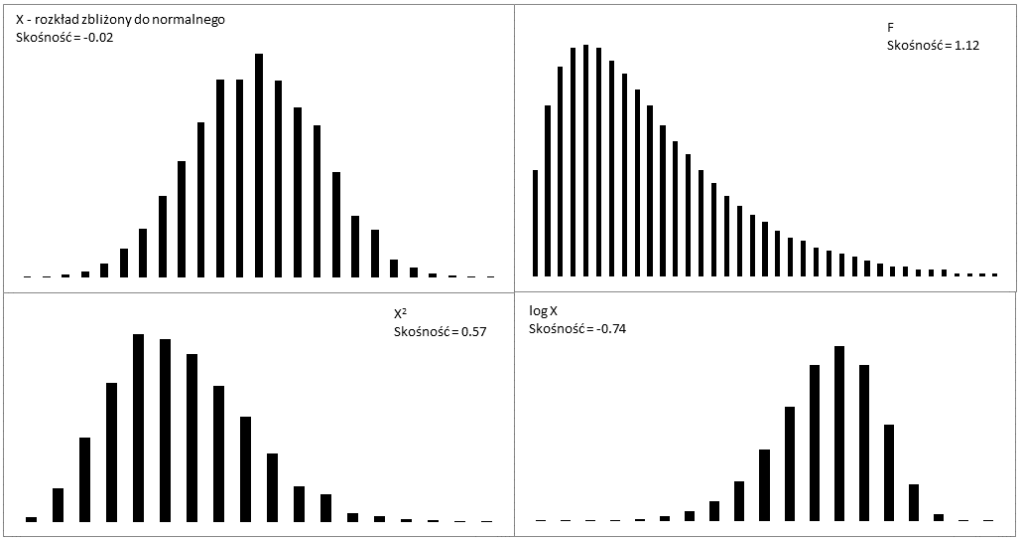
Asymetria skośnego rozkładu powoduje, że znaczenie odchylenia standardowego jest inne od jego znaczenia w rozkładzie symetrycznym, takim jak krzywa normalna lub rozkład t-Studenta (informacje na temat rozkładu t-Studenta znajdziesz w rozdziałach 8. i 9.). W rozkładzie symetrycznym, na przykład normalnym, 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 proporcji 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:

Skośność = (*Średnia* – *Moda*) / *Odchylenie standardowe*

Jednak bardziej typowym sposobem jest użycie sumy podniesionych do sześcianu 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 sześcianów wartości standaryzowanych.

Excel używa odmiany takiej formuły w funkcji `SKOŚNOŚĆ()`:

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

Po odrobinie zastanowienia można zauważyć, że funkcja Excela zawsze zwraca większą wartość niż prosta średnia sześcianów 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 sześcianów wartości *z*. Spójrz na rysunek 7.2, gdzie oryginalne wartości w kolumnie `A` są po prostu powielone w kolumnie `E`. Zauważ, że zarówno średnia sześcianów 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 różnica pomiędzy średnią sześcianów wartości z a wartością zwracaną przez funkcję SKOŚNOŚĆ() jest coraz mniejsza

	A	B	C	D	E	F	G
1	Wartości oryginalne	Wartości standaryzowane (wartości z)	z^3		Wartości oryginalne	Wartości standaryzowane (wartości z)	z^3
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

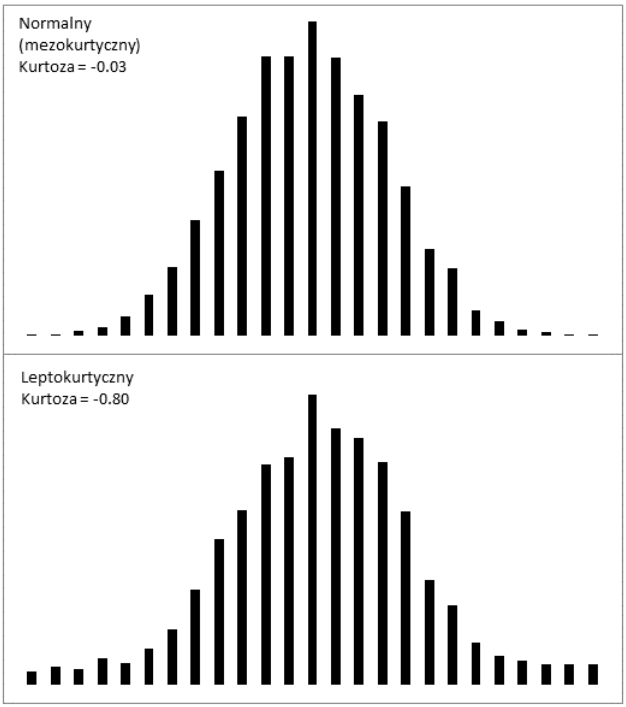
Typy kurtozy

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 mezokurtyczna.
- **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 krzywym 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 załóżmy, 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.

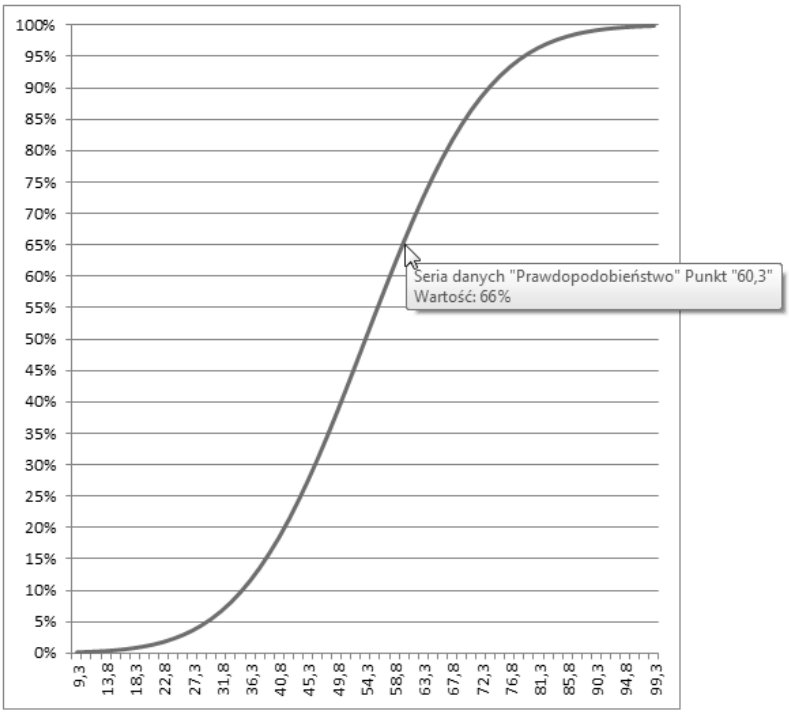
- ***Skumulowany*** — czwarty argument wskazuje, czy chcesz otrzymać skumulowane prawdopodobieństwo poziomów HDL od 0 do *x* (równego w tym przykładzie 60), czy prawdopodobieństwo odpowiadające poziomowi HDL równemu dokładnie *x* (czyli 60). Jeżeli chcesz otrzymać skumulowane prawdopodobieństwo, podaj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

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.

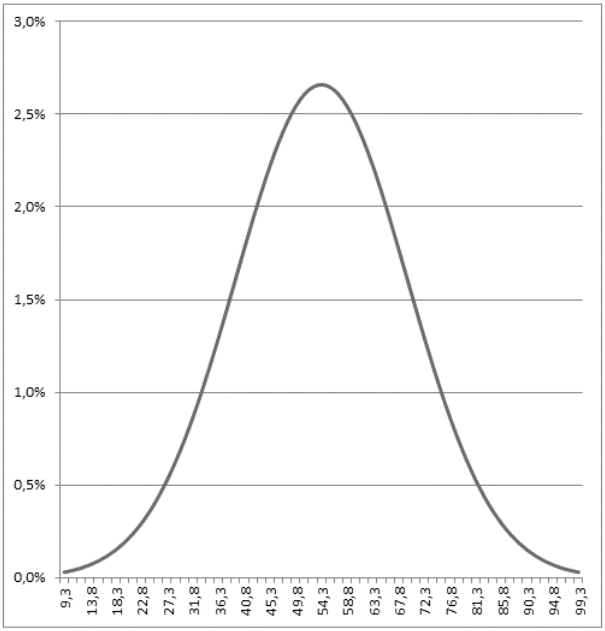


Rysunek 7.4. Możesz dostosować liczbę linii siatki, formatując oś pionową, aby pokazywała więcej lub mniej głównych jednostek

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 (FAŁSZ). 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 zajmiemy. Nie myśl jednak, że możesz używać przedziałów ufności tylko do rozkładów normalnych.

Znaczenie przedziału ufności

Załóżmy, że zmierzyłeś 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 stablicowanych 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, załóżmy, 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:

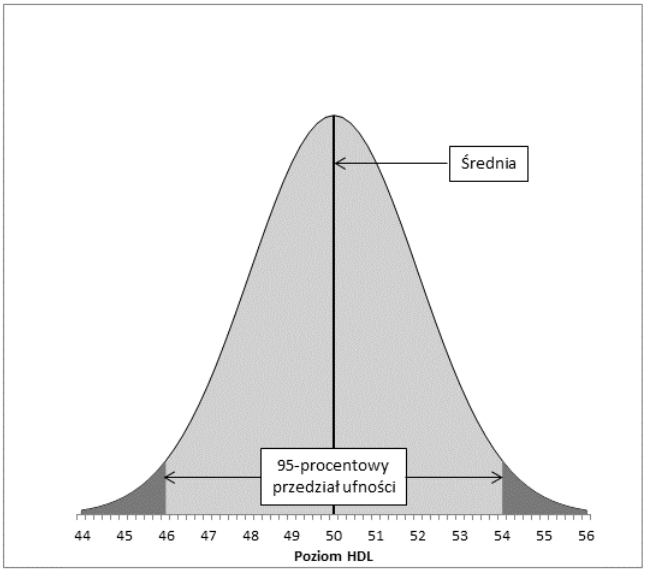
```
=ROZKŁ.NORMALNY.S.ODWR(0,025)
=ROZKŁ.NORMALNY.S.ODWR(0,975)
```

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.

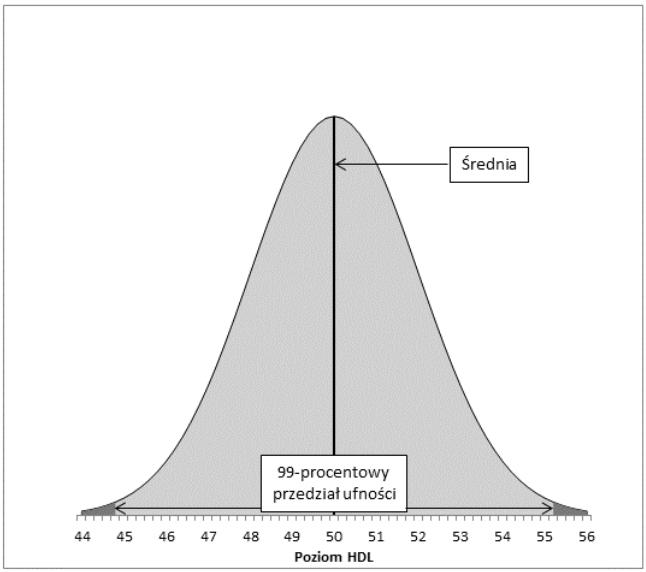
Podobnie funkcja `ROZKŁ.NORMALNY.S.ODWR(0,975)` zwraca 1,96, czyli wartość *z*, która ma 97,5% pola pod krzywą po lewej stronie. Inaczej można powiedzieć, że 2,5% pola pod krzywą leży po jej prawej stronie. Te liczby są pokazane na rysunku 7.6.

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 węższy 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:

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

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:

```
=ROZKŁ.NORMALNY.S.ODWR(0,025)
=ROZKŁ.NORMALNY.S.ODWR(0,975)
```

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:

```
=ROZKŁ.NORMALNY.S.ODWR(0,005)
=ROZKŁ.NORMALNY.S.ODWR(0,995)
```

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 odchyleń standardowych — dlatego wartość *z* równa 1,96 odpowiada punktowi, który znajduje się 1,96 odchyleń standardowych powyżej średniej, a –1,96 znajduje się 1,96 odchyleń 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 =UFNOŚĆ.NORM(F2;C2;ILE.LICZB(A2:A17))

	A	B	C	D	E	F	G	H	I
1	HDL	Średnia HDL	Odchylenie standardowe populacji			Alfa	Połowa szerokości przedziału		
2	88	57,19	22,00			0,05	10,78		
3	64								
4	50				Przedział ufności:		46,41	do	67,97
5	67								
6	45								
7	86						Wartość z		
8	71				Alfa/2	0,025	-1,96		
9	68				1-(Alfa/2)	0,975	1,96		
10	36								
11	20				Przedział ufności:		46,41	do	67,97
12	57								
13	49								
14	37								
15	94								
16	39								
17	44								

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 odpowiednie 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 odchyleń 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 ufność 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 ufności.

Rysunek 7.9. Przy jednakowych pozostałych parametrach przedział ufności zbudowany przy użyciu rozkładu t-Studenta jest szerszy niż skonstruowany za pomocą rozkładu normalnego

G2 =UFNOŚĆ.T(F2;C2;ILE.LICZB(A2:A17))

	A	B	C	D	E	F	G	H	I
1	HDL	Średnia HDL	Odchylenie standardowe próby			Alfa	Połowa szerokości przedziału		
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						**Wartość t**		
8	71				Alfa/2	0,025	-2,13		
9	68				1-(Alfa/2)	0,975	2,13		
10	36								
11	20				Przedział ufności:		46,01	do	68,36
12	57								
13	49								
14	37								
15	94								
16	39								
17	44								

Zauważ, że 95-procentowy przedział ufności na rysunku 7.9 rozciąga się od 46,01 do 68.36, natomiast na rysunku 7.8 od 46,41 do 67,97. Przedział ufności 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 ufności jest szerszy.

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

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.

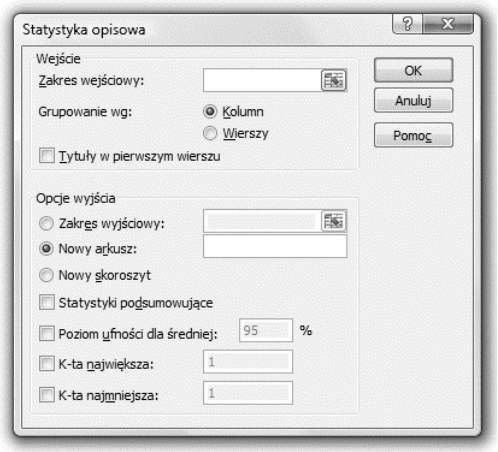
Zastosowanie dodatku Analiza danych do przedziałów ufności

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.

	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			

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

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

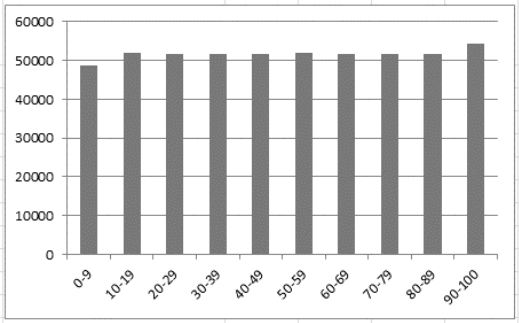
Istnieje zagadnienie łączące średnią i rozkład normalny, o którym dotychczas nie wspominaliśmy. To centralne twierdzenie graniczne — groźnie brzmiące pojęcie o prostym zastosowaniu. Wygląda to tak jak w poniższej baśni.

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 naprawdę 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

	A	B
1	Etykiety wierszy	Liczba pojazdów w okresie
2	0-9	48601
3	10-19	52053
4	20-29	51616
5	30-39	51789
6	40-49	51694
7	50-59	51886
8	60-69	51661
9	70-79	51710
10	80-89	51766
11	90-100	54224
12	Suma końcowa	517000
13		

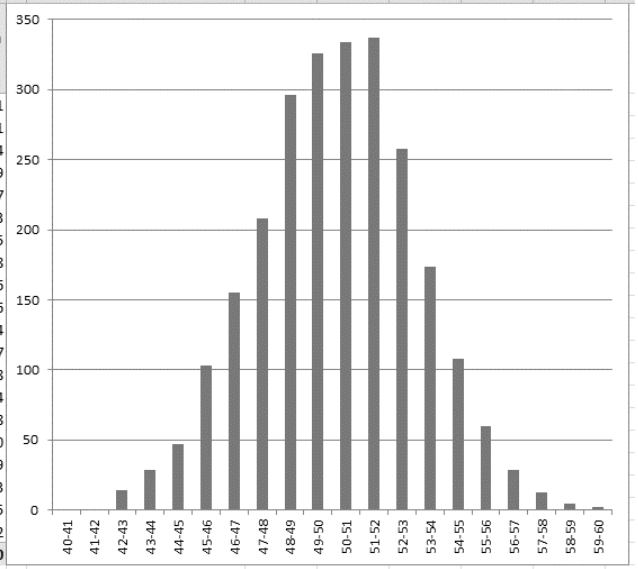


Na rysunku 7.12 różne zakresy pojazdów są pokazane jako „etykiety wierszy" w komórkach `A2:A11`. Dlatego na przykład było tam 48 601 wystąpień z pomiędzy 0 a 9 pojazdami przekraczającymi skrzyżowanie w ciągu dwuminutowych okresów. Twoi zbieracze danych zarejestrowali inne 52 053 wystąpienia pomiędzy 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.

	A	C	D
1	Średnia liczba pojazdów na skrzyżowaniu	Etykiety wierszy	Liczba z Średnia liczba pojazdów na skrzyżowaniu
2	47,82	40-41	1
3	43,73	41-42	1
4	51,32	42-43	14
5	49,82	43-44	29
6	47,69	44-45	47
7	52,41	45-46	103
8	48,87	46-47	155
9	52,40	47-48	208
10	45,93	48-49	296
11	52,82	49-50	326
12	47,88	50-51	334
13	45,77	51-52	337
14	54,01	52-53	258
15	48,23	53-54	174
16	49,53	54-55	108
17	54,47	55-56	60
18	54,84	56-57	29
19	47,82	57-58	13
20	45,95	58-59	5
21	47,29	59-60	2
22	50,71	Suma końcowa	2500



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;WIERSZ(A1:A135))*(0,1^WIERSZ(A1:A135))
*(0,9^(1500-WIERSZ(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 minikomputery, 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 powinieneś 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

E

F

G

H

I

J

K

L

Ł

M

N

O

P

R

S

Ś

T

U

V

W

Z

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 SA
ul. Kościuszki 1c, 44-100 Gliwice
tel.: 32 230 98 63
e-mail: helion@helion.pl
http://helion.pl

ISBN 978-83-246-3668-6

9 788324 636686

Cena: 79,00 zł

Informatyka w najlepszym wydaniu