Helion

ANALIZA STATYSTYCZNA

Microsoft Excel® 2016 PL

que

Conrad Carlberg

Tytuł oryginału: Statistical Analysis: Microsoft Excel 2016

Tłumaczenie: Przemysław Janicki

ISBN: 978-83-283-4467-9

Authorized translation from the English language edition, entitled:
STATISTICAL ANALYSIS: MICROSOFT EXCEL 2016; ISBN 0789759055;
by Conrad Carlberg; published by Pearson Education, Inc, publishing as QUE Publishing.

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/anstae
Możesz tam wpisać swoje uwagi, spostrzeżenia, recenzję.

Pliki z przykładami omawianymi w książce można znaleźć pod adresem:
f*tp://ftp.helion.pl/przyklady/anstae.zip*

Printed in Poland.

Spis treści

Prawdomówność statystyki

Kilkadziesiąt lat temu niejaki Darrell Huff napisał książkę zatytułowaną *How to Lie with Statistics* („Jak kłamać za pomocą statystyki"). W książce opisano różnorodne zdumiewające sposoby użycia statystyki, zastosowane często w sposób niezamierzony i powodujące zmylenie odbiorców.

Podczas przygotowywania tej książki przejrzałem ponownie publikację Huffa (nie byłem nawet w przedszkolu, gdy została ona wydana) i przypomniała mi ona, że częstym powodem zejścia na manowce podczas stosowania statystyki jest błędny kontekst.

W kolejnym rozdziale będziemy kontynuować zapoczątkowane w rozdziale poprzednim przejście od statystyki opisowej do wnioskowania statystycznego — wnioskowania o parametrach i rozkładach w populacji na podstawie ich odpowiedników w próbie. Zanim jednak wejdę głębiej w tematykę wnioskowania statystycznego z wykorzystaniem Excela, powinienem, jak sądzę, zwrócić Twoją uwagę na to, że w pewnych sytuacjach zarówno statystyka opisowa, jak i wnioskowanie statystyczne mogą sprowadzić Cię na manowce.

W moim przekonaniu źródłem problemu są przede wszystkim trzy kwestie:

- pozyskiwanie danych na drodze błędnie zaplanowanego eksperymentu,
- niezrozumienie sposobu działania aplikacji statystycznych lub błędna interpretacja uzyskiwanych w nich wyników,
- utrata kontroli nad warunkami przebiegu eksperymentu.

8

W TYM ROZDZIALE:

Dlatego zamierzam poświęcić część tego rozdziału na omówienie **kontekstu** analizy statystycznej, tzn. jak stworzyć sytuację, w której wyniki analizy faktycznie będą miały przypisywane im znaczenie. Kiedy dane są zbierane poza kontekstem ściśle zdefiniowanego eksperymentu, nie są wiarygodne. Gorzej nawet: jak zauważa Huff, łatwo mogą wprowadzić w błąd. Aby upewnić się na sto procent, że umiejscowiłeś swoje analizy we właściwym kontekście, powinieneś zadać sobie wprost pytanie o to, co może zagrażać wiarygodności Twoich badań. Najlepszą odpowiedzią na to pytanie będzie w pełni zgodne z regułami sztuki zaplanowanie całego eksperymentu.

Większą część tego rozdziału zamierzam poświęcić omówieniu problemów, które wynikają głównie ze sposobu, w jaki Excel implementuje i dokumentuje wybrane narzędzia służące do automatyzacji analiz statystycznych.

W dalszej części rozdziału nawiązuję również do pewnego eksperymentu poznawczego, który być może pozwoli nam odpowiedzieć na pytanie, czy wyniki mogą być zreplikowane (odtworzone). Odpowiedzi tej poszukują obecnie członkowie grupy badawczej z USA — być może słyszałeś o tym eksperymencie, zwanym często **projektem badania replikowalności**.

Zakładam, że nie sięgnąłeś po tę książkę — a w każdym razie nie na tym etapie jej lektury — by dowiedzieć się, w jakich sytuacjach (i dlaczego) statystyka może okazać się zwodnicza. Mogę więc jedynie zachęcić Cię, byś przeczytał ten rozdział z uwagą i wziął sobie do serca przynajmniej część z płynących z niego wniosków. Jeżeli niewłaściwie zaplanujesz swój eksperyment badawczy, dalsza analiza uzyskanych wyników będzie pozbawiona sensu: będzie stratą czasu zarówno dla samego badacza, jak i odbiorców rezultatów jego pracy. Pamiętaj też, że nie ma lepszego sposobu na utratę wiarygodności w oczach środowiska (szczególnie w środowisku badaczy) niż wyrażenie przekonania, że oprogramowanie samo dobrze wie, co robi.

Kontekst wnioskowania statystycznego

Statystyka dostarcza sposobu badania, jak osoby i rzeczy reagują na świat, i jako taka jest fascynująca, irytująca, a czasami kontrowersyjna. Szczególnie statystyka opisowa jest praktykowana z dużym upodobaniem przez niektóre osoby. Kibice sportowi potrafią przerzucać się średnią trafień, dorobkiem napastników oraz osiągnięciami w grze swoich ulubionych graczy.

Pokrewna gałąź statystyki, wnioskowanie statystyczne, wykorzystuje nieco inne podejście, opierające się m.in. na budowie formalnych testów wykorzystujących pomiary średnich arytmetycznych, odchyleń standardowych i korelacji nie tylko do oszacowania parametrów w populacji, ale również do oceny stopnia precyzji tych szacunków.

Właśnie o weryfikacji hipotez myśli większość Czytelników tej książki, gdy napotyka termin *statystyka*. Jest to naturalne, skoro na początku studiów czytali o badaniach wykorzystujących wnioskowanie statystyczne, a później na zajęciach laboratoryjnych konstruowali własne badania, zbierali dane i za pomocą metod wnioskowania statystycznego podsumowywali liczby, aby na ich podstawie móc wyciągnąć bardziej ogólne wnioski.

Można to zrozumieć, ponieważ statystyka na studiach jest po prostu często źle wykładana. Być może Twoje doświadczenie jest inne — mam taką nadzieję — ale wiele osób nigdy nie podejmuje dodatkowej nauki statystyki po spełnieniu wymagań uczelni lub wydziału. Przeżyłem to samo w pewnym cenionym kolegium sztuk wyzwolonych. Statystyka nie interesowała mnie do czasu studiów magisterskich, kiedy to rozpocząłem pobieranie nauk od osób, które rzeczywiście wiedziały, o czym mówią.

Mimo tego statystyka wydaje się pełnić wiodącą rolę w badaniach doświadczalnych w kolegiach i na uniwersytetach. Jest to jednak przerost formy nad treścią. Gdy przychodzi do prowadzenia rzeczywistych badań, okazuje się, że statystyka jest *najmniej* ważnym narzędziem w dostępnej Ci palecie rozwiązań.

Jestem w zasadzie pewien swoich racji. Spędziłem lata, czytając raporty z badań, w których dużo wysiłku włożono w analizę statystyczną. Podczas tych badań nie postarano się jednak zbytnio w fazie projektowania i wdrażania eksperymentu, który nadałby statystyce rzeczywisty sens.

W połowie lat 60. ubiegłego wieku Donald Campbell i Julian Stanley opublikowali monografię zatytułowaną *Experimental and Quasi-Experimental Designs for Research* („Eksperymentalne i quasi-eksperymentalne projekty badań"). To opracowanie, znane szerzej pod nazwiskami autorów, opiera się na rozróżnieniu pomiędzy dwoma typami trafności: uogólnianiem, czyli zewnętrzną trafnością, i wewnętrzną trafnością.

Campbell i Stanley stwierdzili, że oba typy trafności są niezbędne do tego, aby badanie eksperymentalne było przydatne. Musi być poprawne wewnętrznie, czyli zaprojektowane tak, aby jego procedury były wiarygodne.

Równocześnie eksperyment musi być poprawny zewnętrznie, czyli nadawać się do uogólnienia: obiekty badań muszą być wybrane tak, aby dało się uogólnić wyniki doświadczalne na populację, która nas interesuje. Firma farmaceutyczna mogłaby przeprowadzić badanie, które wykazałoby z całkowitą wewnętrzną trafnością, że nowy lek nie ma znaczących efektów ubocznych. Jeżeli jednak przedmiotem badań byłyby mrówki, nie zażywałbym tego leku.

Zapewnienie trafności wewnętrznej

Poprawny eksperyment, tzw. „złoty standard" projektu, zaczyna się od losowego wybrania obiektów z populacji, dla której chcesz przeprowadzić uogólnienia i wyciągać wnioski. (Jeżeli testujesz lek z myślą o populacji wszystkich ludzi, nie powinieneś ograniczać

składu próbki jedynie do studentów). Następnie przyjmujesz α, czyli poziom błędu: ryzyko, które jesteś w stanie ponieść, wnioskując *błędnie*, że Twoje działanie (np. kuracja lecznicza) przyniosło efekt.

UWAGA

Istnieje kilka wspaniałych opracowań na temat budowania dobrych planów próbkowania. Są to m.in. *Sampling Techniques* („Techniki próbkowania") Williama Cochrana (1977) i *Survey Sampling* („Próbkowanie badań") Lesliego Kisha (1995).

Następnym krokiem jest losowe przypisanie obiektów do jednej z dwóch lub wielu grup. W najprostszym projekcie występuje jedna grupa eksperymentalna i jedna „kontrolna" lub „porównawcza". Traktujesz swoją grupę eksperymentalną w określony sposób i aplikujesz odmienne traktowanie grupie porównawczej lub po prostu pozostawiasz ją samą sobie. W końcu wykonujesz pewien rodzaj pomiarów związanych z eksperymentem: jeżeli podawałeś statynę, możesz zmierzyć poziom cholesterolu badanych osób. Jeżeli pokazałeś jednej z grup podburzający blog polityczny, możesz zapytać jej członków o zdanie na temat pewnego polityka. Jeżeli zastosowałbyś różne rodzaje nawozów sztucznych do różnych zbiorów uprawianych drzew cytrusowych, mógłbyś zobaczyć, jak ich owoce różnią się miesiąc później.

W końcu mógłbyś przepuścić wyniki pomiarów przez taką czy inną procedurę statystyczną, aby zobaczyć, czy np. dane uzyskane na drodze eksperymentu przeczą hipotezie braku efektu oddziaływania danego czynnika przy zadanym poziomie błędu (α).

Głównym przesłaniem tej nieco jeszcze chaotycznej opowieści jest to, że potrzebne są dwie grupy, które są równoważne we wszystkich aspektach oprócz jednego — efektu oddziaływania na jedną z grup, którego nie doświadczyła druga grupa. Odbywający się na początku losowy przydział do grup pomaga zmniejszyć ryzyko wystąpienia systematycznej różnicy pomiędzy grupami. Takie samo traktowanie obydwu grup, z wyjątkiem wpływu badanego czynnika, pomaga wyizolować ten czynnik jako jedyne źródło zaobserwowanych różnic. Te właśnie różnice mamy zmierzyć w ramach eksperymentu.

Jeżeli sposób, w jaki wyodrębniłeś obie grupy, uwiarygadnia stwierdzenie, że jedyna znacząca różnica pomiędzy nimi jest wywołana oddziaływaniem danego czynnika, Twój eksperyment jest określany jako wewnętrznie trafny. Wewnętrzne porównanie pomiędzy dwiema grupami jest prawidłowe.

Jeżeli obiekty badań są reprezentatywne dla populacji, na którą chcesz uogólnić swoje wnioski, eksperyment jest określany jako spełniający wymogi trafności zewnętrznej. Wtedy uogólnienie wniosków z próby losowej na populację jest uzasadnione.

Zagrożenia trafności wewnętrznej

Oprócz błędu próbkowania Campbell i Stanley zidentyfikowali i opisali około siedmiu zagrożeń mogących podważyć wewnętrzną trafność eksperymentu. Ustanowienie przez losowy dobór (i odpowiednie późniejsze zarządzanie) równoważnych grup eksperymentalnej i kontrolnej służy do eliminacji większości z tych zagrożeń.

Dobór

Sposoby doboru obiektów do grup eksperymentalnej i kontrolnej mogą zagrozić wewnętrznej spójności doświadczenia, szczególnie jeżeli poszczególne obiekty współdecydują o swoim statusie. Załóżmy, że badacz chciałby porównać proporcje sukcesów dwóch procedur medycznych, z których każda jest prowadzona w innym szpitalu w dużym mieście.

Podczas porównywania wyników tych dwóch procedur nie jest możliwe wyznaczenie, czy dowolna różnica — powiedzmy, w proporcjach przeżycia — wynika z kuracji, czy z różnic w populacjach, z których pochodzą pacjenci. Może to być niewykonalne, ale zwykle zaleca się przypisywanie uczestników losowo do badanych grup, co zgodnie z oczekiwaniem ma wyrównać efekt przynależności do jednej lub drugiej populacji. Obciążenie wyboru w badaniach przeprowadzanych na dużą skalę może być kontrolowane dzięki zbieraniu wyników z wielu szpitali, którym losowo przydzielono stosowanie jednej lub drugiej kuracji. (To podejście może powodować inne problemy).

Historia

Może wystąpić zdarzenie o dużej sile oddziaływania, które będzie mieć wpływ na reakcję obiektów na działanie eksperymentu. Załóżmy, że badasz efekt wpływu kampanii politycznej na nastawienie elektoratu do osoby rządzącej. W tym samym czasie ma miejsce krach finansowy, który dotyka wszystkich, bez względu na preferencje polityczne. Odróżnienie efektu kampanii od efektu kryzysu jest teraz bardzo trudne. Jednak przy założeniu, że jego wpływ na obie grupy (odbiorców kampanii oraz grupę kontrolną złożoną z osób, które nie miały okazji śledzić jej przebiegu) jest w przybliżeniu jednakowy, możesz przypisać zaobserwowane różnice efektowi kampanii. Bez równoważnych grup, eksperymentalnej i kontrolnej, odróżnienie efektów kampanii od wpływu czynników zewnętrznych byłoby niemożliwe.

Inny przykład: jeżeli osoby badające wpływ kuracji medycznej są świadome, który z pacjentów znajduje się w której grupie, jest możliwe, że ich zachowanie może zakłócić efekty oddziaływania, gdy (zwykle niechcąco) będą sygnalizowali swoje oczekiwania względem poszczególnych grup lub też będą delikatnie wpływać na zachowanie pacjentów, tak by było ono zgodne z oczekiwanymi wynikami. Aby temu zapobiec — aby świadomość przynależności do grupy, a zatem odmiennego traktowania, nie miała wpływu na wyniki — zwykle stosuje się metodę podwójnie ślepej próby, która oznacza, że zarówno osoba nadzorująca kurację, jak i ta, której ona dotyczy, nie wiedzą, który specyfik, lek czy placebo, jest podawany danemu pacjentowi.

Narzędzia

Stosowany tutaj termin *narzędzia* wykracza poza urządzenia pomiarowe, takie jak suwmiarka, i obejmuje wszystkie metody, które zwracają informacje ilościowe, takie jak choćby prosty kwestionariusz. Zmiana w sposobie mierzenia wyników może spowodować duże trudności w interpretacji. Na przykład, abstrahując od pytania o same porównania między grupą badawczą i kontrolną, wielu badaczy autyzmu wierzyło, że widoczne zwiększenie częstotliwości występowania tej choroby w ostatnich kilku dekadach jest związane głównie ze zmianami w jej diagnozowaniu, które doprowadziły do znacznie częstszej wykrywalności.

Testowanie

Wielokrotne poddawanie obiektów z danej grupy wpływowi badanego czynnika może wpływać na uzyskiwane wyniki. Do pewnego stopnia może ono wzmocnić (lub osłabić) reakcję obiektu na dany czynnik.

Na ten efekt podatni są nie tylko ludzie i inne istoty żywe. Również na przykład metale, które są obiektami wielokrotnie powtarzanych testów obciążeniowych, mogą pod ich wpływem zmienić swoje charakterystyki fizyczne. Mimo wszystko testowanie pozostaje nieodłączną częścią każdego badania ilościowego.

Dojrzewanie

Stopień dojrzałości różni się pomiędzy różnymi przedziałami wieku, co może zmniejszyć wiarygodność niektórych porównań. Nawet gdyby grupa eksperymentalna i kontrolna były równoważne pod względem wieku obiektów dzięki losowemu doborowi i analizie kowariancji (patrz rozdział 17., „Analiza kowariancji — podstawy", i 18., „Analiza kowariancji — dalsze zagadnienia"), może się zdarzyć i tak, że zróżnicowanie obiektów, które wystąpi w trakcie badania, będzie skutkowało tym, że rozróżnienie pomiędzy efektami wynikającymi z oddziaływania badanego czynnika a skutkami tego zjawiska stanie się trudniejsze.

Regresja

Regresja w stronę średniej (patrz rozdział 4., „Jak zmienne wspólnie się zmieniają — korelacja") może mieć wyraźny wpływ na wyniki eksperymentu, szczególnie gdy obiekty są wybrane *z powodu* ekstremalnych wartości pewnej miary związanej z tą użytą do pomiaru wyników. Będą one podążać w stronę średniej bez względu na jakikolwiek efekt oddziaływania. Użycie dopasowanych par, z których pojedyncze elementy zostaną przypisane do różnych grup, ma na celu większą od zwykłego losowania efektywność wyrównania dwóch grup przed poddaniem ich oddziaływaniu czynnika. Jednak często się zdarza, że efekt regresji niweczy te zamierzenia z powodu niedoskonałej korelacji wyników mierzonych w parach.

Śmiertelność

Śmiertelność podczas eksperymentu dotyczy sytuacji, gdy obiekty z grupy eksperymentalnej lub kontrolnej nie są w stanie dotrwać do końca swojego uczestnictwa w eksperymencie. (W tym kontekście „śmiertelność" nie musi oznaczać straty uczestników z powodu ich śmierci — dotyczy dowolnego efektu lub efektów, które powodują, że obiekty przestają uczestniczyć w eksperymencie). Chociaż losowe przypisanie na początku pomaga wyrównać grupy pod względem prawdopodobieństwa późniejszej utraty obiektów, bardzo trudno odróżnić opuszczenie próby z przyczyn leżących po stronie samego eksperymentu od opuszczenia jej z innego powodu. Ten problem jest szczególnie wyraźny w badaniach medycznych, gdzie w wielu eksperymentach biorą udział osoby, których przewidywana długość życia jest względnie krótka.

Przypadek

Pod koniec eksperymentu, gdy wszystko przeprowadzono zgodnie z regułami sztuki, wykonano niezbędne badania i pomiary itp., na scenę wchodzi analiza statystyczna. Zwykle używasz jej do oceny tego, jakie jest prawdopodobieństwo, że uzyskane wyniki otrzymałeś przez czysty przypadek, tak że rezultaty podobnego badania na całej populacji byłyby inne, gdybyś tylko mógł je przeprowadzić.

Jeżeli postąpisz zgodnie z tzw. „złotym standardem" losowego doboru, zrobisz wszystko, co tylko możliwe, aby ustanowić równoważne grupy — grupy, które mają następujące właściwości:

- Nie są wynikiem autodoboru obiektów ani żadnego rodzaju systematycznego przypisania, które mogłoby wprowadzić wcześniej istniejące obciążenie.
- Podlegają tym samym historycznym zdarzeniom podczas eksperymentu — od politycznych zamieszek do przypadkowego wprowadzenia kurzu do wrażliwego środowiska produkcyjnego.
- Są mierzone tym samym zestawem narzędzi w trakcie całego przebiegu eksperymentu.
- Prowadzący test nie dają grupom odczuć różnego traktowania.
- Członkowie wszystkich grup dojrzewają w równym stopniu w trakcie trwania eksperymentu.
- Obiekty testów nie zostały przypisane do grup na podstawie ekstremalnych wartości pewnych cech.
- Obiekty nie opuszczały każdej z grup z różnym nasileniem.

Losowy wybór i przypisanie obiektów stanowią łącznie najlepszy sposób zapewnienia, że grupy eksperymentalne posiądą wymienione cechy. Jednak i te techniki nie są doskonałe. Mimo wszystko może zdarzyć się i tak, że pewne czynniki zewnętrzne będą wywierały większy wpływ na jedną grupę niż na inną lub że dobór losowy nie wyeliminował efektu wstępnego obciążenia grup albo też że więcej niż tylko przypadek wpłynął na różną śmiertelność obiektów w poszczególnych grupach itd.

Zatem czynniki mogące podważyć założenie o trafności wewnętrznej eksperymentu istnieją i chociaż będziesz robił, co tylko w Twojej mocy, by je zmniejszyć, nigdy nie możesz całkowicie wykluczyć ich wpływu jako alternatywnego wyjaśnienia obserwowanych wyników.

W zależności od stopnia nasilenia tych zagrożeń analiza statystyczna może stracić swój sens. Tradycyjnie używana podczas testowania hipotez analiza statystyczna służy do określenia ilościowego roli przypadku w wyniku eksperymentu. Jednak dokładne przypisanie stopnia, w jakim przypadek odgrywa swoją rolę, zależy od obecności dwóch lub więcej grup, które są sobie równoważne z jednym jedynym wyjątkiem — aplikacji (lub jej braku) wpływu badanego czynnika.

Rozważmy taką sytuację: przez jeden miesiąc podawałeś nowy lek grupie eksperymentalnej, a grupie kontrolnej w jego miejsce poddawałeś przez ten czas placebo. Lek ma na celu redukcję poziomu tzw. złego cholesterolu, czyli lipoprotein niskiej gęstości (LDL) w krwi. Na koniec miesiąca próbki krwi zostały pobrane i przeprowadziłeś analizę statystyczną wyników. Ta analiza pokazała, że prawdopodobieństwo tego, iż średni poziom LDL w grupie eksperymentalnej i kontrolnej pochodzą z tej samej populacji, wynosi ok. 1 do 1000.

Wniosek, że średnie grup pochodzą z tej samej populacji, oznaczałby, iż kuracja nie doprowadziłaby do powstania populacji, których średnie poziomy LDL różniłyby się w stopniu usprawiedliwionym przez zażywanie leku. Jednak wynik analizy uznaje za statystycznie istotny wniosek, że grupy pochodzą teraz z dwóch różnych populacji. Wydaje się to wspaniałą wiadomością… **jeżeli** starannie porównałeś te grupy na początku i utrzymałeś ten poziom równoważności. W przeciwnym przypadku nie możesz stwierdzić, że różnica wynika ze stosowania leku. Mogłaby powstać na przykład dlatego, że członkowie grupy kontrolnej zaprzyjaźnili się i co dzień po przyjęciu swoich placebo chodzili na cheeseburgery.

Możemy wyobrazić sobie łatwo sytuacje, kiedy analiza statystyczna nie zostaje poprzedzona eksperymentem w pełnym tego słowa znaczeniu. Na przykład tworzenie i analiza testów psychologicznych oraz ankiet politycznych są uzupełniane analizą regresji (która jest podstawą większości analiz opisanych w drugiej części tej książki). Nie muszą być one jednak w żadnym razie ograniczone do testowania możliwości poznawczych osobników czy badania ich przekonań politycznych, ale mogą obejmować inne obszary — od medycyny i testowania leków do kontroli jakości w środowiskach produkcyjnych. Ich budowa i interpretacja zależy w dużej mierze od wybranych metod analizy statystycznej — które ta książka omawia (przy zastosowaniu Excela jako platformy obliczeniowej). Nie zmienia to jednak faktu, że nie formułują one żadnej hipotezy, która byłaby poddawana weryfikacji. Mają one bardziej za zadanie ocenę samego testu i tego, co jest przedmiotem badania, w różnych grupach.

Tym niemniej stosowanie analizy statystycznej do wyeliminowania przypadku jako czynnika wpływającego na wyniki eksperymentów jest czymś normalnym, typowym i powszechnym. Gdy myślimy o wynikach eksperymentu, uwzględniając warunki, sytuację, a nawet chorobę, która nas interesuje, chcemy znać naturę użytej analizy statystycznej. W przypadku doświadczeń analiza statystyczna jest *bezcelowa*, jeżeli nie zostanie wykonana w kontekście solidnego projektu eksperymentalnego — takiego, który jest dokładnie zaplanowany i prowadzony.

Problemy z dokumentacją Excela

Jednym z głównych założeń (a zarazem przesłań) tej książki jest to, że Excel *jest* dokładnym i rzetelnym narzędziem do wykonywania analiz statystycznych. Około dwudziestoletnia praktyka z korzystaniem z funkcjonalności Excela — łącznie z zaglądaniem do wnętrza kodu, by przekonać się, jak one działają — upewnia mnie o słuszności tego stanowiska.

Nie oznacza to jednak, że każda funkcja Excela jest bezwarunkowo warta pieniędzy, które wydałeś na jego zakup. Nawet gdy ograniczymy się jedynie do funkcji arkusza — podstawowych elementów składowych każdej analizy z użyciem tego narzędzia — nie da się ukryć, że uzyskanie wiarygodnych rezultatów może wymagać od Ciebie wiele ostrożności i uwagi. W rozdziale 15., „Analiza regresji wielorakiej i rekodowanie zmiennych nominalnych — podstawy", wspominam na przykład o tym, że jedna z podstawowych funkcji statystycznych Excela przez wiele lat zawierała błąd, który sprawiał, że potrafiła zwrócić sumę kwadratów reszt, która była ujemna.

W poprzednim podrozdziale skupiliśmy się na omówieniu tego, jak nie do końca przemyślany proces doboru danych może sprawić, że dalsza analiza (i uzyskane wyniki) będą pozbawione sensu. Większość pozostałych pomyłek podczas przeprowadzania analizy statystycznej jest związana z błędnym zrozumieniem podstawowych pojęć. Niestety, Excel gdzieniegdzie daje nam tego rodzaju mylące wskazówki. Znajdują się one głównie w dodatku, który towarzyszy Excelowi od połowy lat 90. ubiegłego wieku. Był on znany jako *Analysis ToolPak* lub *ATP*, a ostatnio jako dodatek *Analiza danych*. Dodatek ten jest zbiorem narzędzi statystycznych. Ma na celu dostarczenie użytkownikowi sposobów przeprowadzania (przede wszystkim) analiz wnioskowania statystycznego, takich jak analizy wariancji i regresji. Takie analizy możesz przeprowadzić bezpośrednio w arkuszu za pomocą funkcji Excela. Jednak narzędzia dostępne w tym dodatku porządkują i prezentują te analizy poprzez zastosowanie praktycznych formatów i opcji okien dialogowych zamiast dość niewygodnych argumentów funkcji. Dlatego mogą one ułatwić życie. Ponieważ dodatek ten jest dostarczany wraz z Excelem, wielu nowych użytkowników może zakładać (całkiem racjonalnie), że grupuje on wszystkie funkcjonalności Excela przydatne do analiz statystycznych. W rzeczywistości zawiera on trzy metody analizy wariancji, testy *z* i testy *t*, narzędzia do wyznaczania macierzy kowariancji i analizy korelacji, statystyki opisowe itd.

Jednak narzędzia mogą także zmylić lub po prostu nie poinformować o konsekwencjach podjęcia pewnych decyzji. Oczywiście ta uwaga może dotyczyć dowolnego oprogramowania statystycznego, ale szczególnie narażeni są użytkownicy dodatku *Analiza danych*, ponieważ jego dokumentacja jest niezwykle skąpa.

Dobrym przykładem jest narzędzie *Wygładzanie wykładnicze*. Ogólnie rzecz biorąc, jest to metoda wykorzystująca średnie ruchome do prognozowania kolejnej wartości w szeregu czasowym. Zwracane przez nią wyniki zależą w dużym stopniu od parametru liczbowego, tzw. **stałej wygładzania**. Służy ona do uwzględniania w prognozie efektu błędów popełnianych w poprzednich okresach prognozy, a więc ułatwia autokorektę.

Co z tego, skoro sam wybór wartości stałej wygładzania nie jest prosty: wymaga podjęcia decyzji odnośnie do charakteru szeregu wynikowego (czy zależy nam na mniejszym błędzie, czy na większym „wygładzeniu szeregu wyjściowego), identyfikacji występujących w danych trendów (rosnący, malejący, brak trendu) itp. Jakby tego było mało, to choć stosując tę metodę, podajemy zwykle wartość stałej wygładzania, i tak narzędzie wbudowane w dodatek do Excela z niewiadomych przyczyn domaga się podania tzw. **współczynnika tłumienia**. Pojęcie *stałej wygładzania* pojawia się w różnego rodzaju publikacjach mniej więcej dziesięć razy częściej niż *współczynnik tłumienia*, zatem nie ma racjonalnego powodu, by od nowego użytkownika tego narzędzia oczekiwać, że będzie wiedział, czym on jest. Współczynnik tłumienia jest dopełnieniem do 1 stałej wygładzania, więc w praktyce problem ten okazuje się trywialny, ale jest dobrym przykładem niepotrzebnego zamieszania, jakie spowodowali twórcy Excela. Dobra dokumentacja, pisana z myślą o użytkownikach, powinna posługiwać się raczej terminem, który jest szerzej znany (stała wygładzania), albo przynajmniej wyjaśniać, w jaki sposób można uzyskać wartość wymaganą w narzędziu. Dokumentacja stworzona przez Microsoft przez wiele lat nie spełniała tego warunku. Firma zmieniła jej treść dopiero wraz z wydaniem Excela 2013, ale — niestety — choć nazywa go teraz stałą wygładzania, nadal oczekuje podania współczynnika tłumienia. I bądź tu mądry!

Podobnie ma się sytuacja z innymi narzędziami dostępnymi w dodatku *Analiza danych*, a sprawę pogarsza jeszcze fakt, że zwracane przez nie wyniki nie mają postaci formuł, ale wartości liczbowych. To wszystko znacznie utrudnia próbę zrozumienia, co dane narzędzie faktycznie liczy i zwraca jako wynik.

Załóżmy na przykład, że jedno z narzędzi dodatku *Analiza danych* informuje Cię, że średnia wartość konkretnej zmiennej wynosi 4,5 — dodatek umieszcza w komórce wartość 4,5. Jeżeli ta wartość wyda Ci się nieprawidłowa, znalezienie źródła rozbieżności będzie wymagać żmudnej pracy. Natomiast gdyby dodatek pokazywał *formułę* kryjącą się za wynikiem 4,5, byłbyś już na drodze do znacznie szybszego rozwiązania problemu.

Parę narzędzi jest dość dobrych: *Korelacja* i *Kowariancja* dostarczają wyniki, których obliczenie za pomocą wbudowanych funkcji arkusza byłoby żmudne, nie mylą się ani nie są szczególnie trudne do opanowania, a ich wyniki są przydatne w sensie praktycznym. Ale są to wyjątki. (Nadal jednak dostarczają wyniki w postaci stałych wartości zamiast formuł, co jest niewygodne). W celu wyczerpującego przedstawienia przykładu problemów, jakie opisuję, w drugiej części tego rozdziału zajmę się dość szczegółowym omówieniem jednego z narzędzi — *Test F: z dwiema próbami dla wariancji*. Mam ku temu dwa powody:

- Swego czasu statystycy przeprowadzali ten test, aby upewnić się co do braku różnic w wartościach średnich między dwoma populacjami. Z czasem wykazano jednak, że efekt naruszenia tego założenia jest w wielu sytuacjach nieistotny. Nadal istnieją dobre powody korzystania z tego narzędzia, szczególnie w zastosowaniach produkcyjnych, które zależą od analiz statystycznych, jednak prawie nie ma dokumentacji na temat tej techniki, a zwłaszcza dotyczącej posługiwania się nią w dodatku *Analiza danych* Excela.

- Rozważanie problemów z dodatkiem pozwala dobrze poznać sprawy, o które musisz się zawsze zatroszczyć, gdy zaczynasz używać nieznanego oprogramowania statystycznego — również Excela. Jeżeli coś Cię zadziwia, nie bierz nic na wiarę. Zbadaj to dogłębnie.

Mam nadzieję, że dzięki tej prezentacji zrozumiesz, że i drugi kraniec statystycznego continuum — etap czystej analizy, następujący po zaplanowaniu eksperymentu — również wymaga Twojej uwagi. I że tę uwagę musisz poświęcić nie tylko pracy z dodatkiem *Analiza danych*, ale także wtedy, kiedy pracujesz ze zwykłymi funkcjami arkusza.

Test F z dwiema próbami dla wariancji

Teraz, gdy poświęciłem kilka pierwszych stron tego rozdziału na omówienie, czemu rozumienie statystyki może mieć niższy priorytet — przynajmniej w porównaniu z etapem projektowania eksperymentu — chcę przyjąć inny punkt widzenia i przyjrzeć się temu, dlaczego rozumienie statystyki jest ważne: jeżeli nie rozumiesz założeń i stosowanych metod, prawdopodobnie nie będziesz w stanie interpretować wyników analizy. Innymi słowy, kiedy już dane zostały pozyskane w sposób prawidłowy, analiza liczb *ma* znaczenie. Czasami zdarza się, że oprogramowanie prawidłowo przelicza liczby, ale kiepsko wyjaśnia, co zostało zrobione. Oczekujemy, że dokumentacja oprogramowania dostarczy wyjaśnienia, lecz często jesteśmy rozczarowani. Jedno z narzędzi dodatku *Analiza danych* — *Test F: z dwiema próbami dla wariancji*, jest wspaniałym przykładem, dlaczego dosłowne traktowanie dokumentacji jest złym pomysłem.

UWAGA

Nie chciałbym, byś odniósł wrażenie, że testy *F* w dodatku *Analiza danych* są jakoś szczególnie problematyczne. Ich jakość jest nawet wyższa od średniego poziomu dla pozostałych narzędzi dostarczanych w ramach tego pakietu. Po prostu testy *F* stanowią bardzo dobry punkt wyjścia do dalszego omówienia. Wszystkie uwagi, które formułuję w tej sekcji, reprezentują *ten* sposób myślenia, który chciałbym, by towarzyszył Ci również wtedy, gdy będziesz pracował z innymi pakietami oprogramowania do analiz statystycznych.

Oto treść tej dokumentacji z Pomocy programu Excel 2016:

> Narzędzie oblicza wartość *f* statystyki *F* (lub współczynnika *F*). Wartość *f* zbliżona do 1 stanowi dowód na to, że wariancje rozkładu podstawowego są równe. Jeżeli w tabeli wyników *f* < *1*, „P(F <= f) jednostronna" daje prawdopodobieństwo obserwowania wartości statystyki *F* mniejszej niż *f*, gdy wariancje rozkładu są równe i „Wartość krytyczna jednostronna F" jest wartością krytyczną mniejszą od 1 dla wybranego poziomu istotności *Alfa*. Jeżeli *f* > *1*, „P(F <= f) jednostronna" daje prawdopodobieństwo obserwowania wartości statystyki *F* większej od *f*, gdy wariancje rozkładu są równe i „Wartość krytyczna jednostronna F" daje wartość krytyczną większą od 1 dla wartości *Alfa*.

Rozumiesz? Ja też nie.

Poza innymi zastosowaniami test *F* — koncepcja statystyczna, a nie narzędzie Excela — pomaga wyznaczyć, czy wariancje dwóch różnych prób losowych są równe w populacjach, z których zostały pobrane. Narzędzie *Test F* próbuje wykonać ten test za Ciebie. Jednak, jak widzisz, wymaga to większej znajomości podstaw teoretycznych niż tylko użycia narzędzia do wyprowadzenia przydatnych informacji.

Po co przeprowadzać ten test?

Jak przeczytasz w rozdziałach 10., „Testowanie różnic pomiędzy średnimi — dalsze zagadnienia", i 11., „Testowanie różnic pomiędzy średnimi — analiza wariancji", jednym z podstawowych założeń pewnych testów statystycznych jest to, że różne grupy mają tę samą wariancję — lub równoważnie: to samo odchylenie standardowe — mierzonych wyników. W pierwszej połowie ubiegłego wieku podręczniki zalecały przeprowadzenie testu *F* dla równych wariancji przed weryfikacją hipotezy, że różne grupy mają równe średnie. Jeżeli test *F* wykazywał, że grupy mają różne wariancje, zalecano zaniechanie testowania różnic pomiędzy średnimi, ponieważ nie było spełnione podstawowe założenie takiego testu.

Wtedy nadeszły badania nad „odpornością" testów z lat 50. i 60. Te prace badały skutki naruszania podstawowych założeń wielu testów statystycznych. Statystycy, którzy studiowali te problemy, byli zainteresowani określeniem, czy założenia użyte do wyprowadzenia modeli teoretycznych są istotne podczas rzeczywistego stosowania tych modeli.

Zgodnie z oczekiwaniami niektóre z tych założeń są istotne. Na przykład zwykle ważne jest, aby obserwacje były niezależne od siebie, żeby na przykład wynik testu Jana nie miał wpływu na wynik Janiny, jak mogłoby się to zdarzyć, gdyby Jan i Janina byli bliźniętami, które zostały poddane pomiarowi pewnej cechy biochemicznej. Jeżeli nie korzystasz z testu, który dopuszcza (i odpowiednio kwantyfikuje) ten stopień zależności, wiarygodność wyniku może być wątpliwa.

Jednak założenie o równych wariancjach jest często nieistotne. Gdy wszystkie grupy mają tę samą liczbę obserwacji, ich wariancje mogą się bardzo różnić, co jednak pozostanie bez wpływu na poprawność testu statystycznego. Natomiast kombinacja różnych liczebności grup z różnymi wariancjami może powodować problemy. Załóżmy, że jedna grupa ma 20 obserwacji i wariancję 5, a inna grupa 10 obserwacji i wariancję 2,5. A zatem jedna grupa jest dwukrotnie większa niż druga **i** jej wariancja jest dwukrotnie większa niż drugiej. W tej sytuacji tabele statystyczne i funkcje mogą wskazywać, że prawdopodobieństwo błędnej decyzji wynosi 5%, gdy rzeczywiście jest równe 3%. Jest to dość mała różnica w przypadku tak niezgodnych liczebności prób losowych i wariancji. Dlatego statystycy zwykle uważają niektóre testy za *odporne* pod względem naruszania założenia równych wariancji.

Nie oznacza to, że nie powinieneś używać testu *F* do zdecydowania, czy *wariancje* dwóch prób losowych w populacji są równe. Jeżeli jednak celem jest testowanie różnic *średnich* grup, a liczebności grup są z grubsza równe, wtedy nie powinieneś się tym przejmować. Ponadto jeżeli zarówno liczebności grup, jak i wariancje są bardzo rozbieżne, lepiej spędzić czas, ustalając, dlaczego losowy wybór i losowe przypisanie dały w wyniku te rozbieżności. Zawsze ważniejsze jest upewnienie się, że zaprojektowałeś poprawne porównania, niż skrupulatne wykonanie testu, który może okazać się nadmiarowy.

Przy braku powyższej przesłanki — wykorzystania testu *F* jako wstępu do testowania równości średnich w grupach — zasadność przeprowadzania testu *F* tylko w celu porównania wariancji jest raczej ograniczona. Oczywiście pewne dyscypliny, takie jak badania operacyjne i teoria sterowania, często testują rozproszenia miar jakości. Jednak inne obszary zastosowań, takie jak medycyna, biznes i nauki behawioralne, skupiają się częściej na różnicach wartości średnich niż różnicach miar rozproszenia.

UWAGA

Łatwo pomylić test *F* opisany tutaj z testem *F* używanym w analizie wariancji i kowariancji opisanym w rozdziałach od 10. do 15. Test *F* jest **zawsze** oparty na ilorazie dwóch wariancji. W tym zastosowaniu służy do odpowiedzi na pytanie, czy dwie próbkowane grupy mają różne wariancje populacji, z których pochodzą. W przypadku analizy wariancji i kowariancji skupia się na rozproszeniu średnich grup podzielonym przez rozproszenie wartości w tych grupach. W obydwu przypadkach statystyką testową jest *F* — iloraz wariancji. W obydwu przypadkach porównujesz iloraz *F* z wartością z rozkładu, który jest prawie tak dobrze znany jak rozkład normalny. Jedynie cel tych testów się różni: testowanie różnic w wariancjach jako sam wynik końcowy, a testowanie różnic w wariancjach w celu wnioskowania na temat różnic wartości średnich.

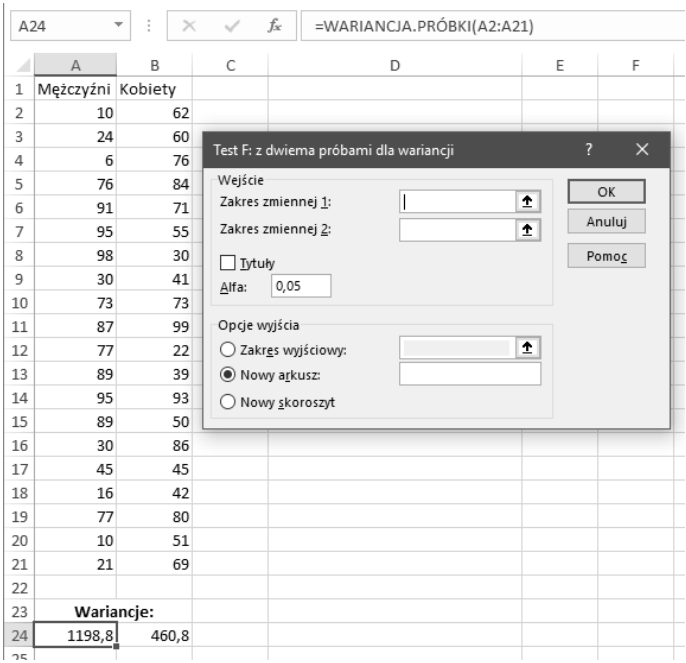
Zakładam, że o ile nie działasz w branży wytwórczej, tylko okazjonalnie będziesz używać narzędzia *Test F: z dwiema próbami dla wariancji*. W takim przypadku chciałbyś wiedzieć, jak ochronić się przed sytuacjami, które mogą Cię zmylić. Jeśli nie, może zechcesz dowiedzieć się bardziej szczegółowo, jak dokumentacja Excela potrafi sprowadzić Cię na manowce.

Korzystanie z narzędzia — przykład liczbowy

Rysunek 8.1 pokazuje przykład, jak możesz użyć narzędzia *Test F*.

Załóżmy, że określiłeś zakres `A1:A21` (*Mężczyźni*) jako *Zakres zmiennej 1* w oknie dialogowym, a `B1:B21` (*Kobiety*) jako *Zakres zmiennej 2*. Zaznaczyłeś pole wyboru *Tytuły*, zaakceptowałeś domyślną wartość *Alfa* równą `0,05` i zaznaczyłeś komórkę `D2` jako lokalizację dla zakresu wyników.

Po kliknięciu *OK* narzędzie *Test F* działa i wyświetla wyniki pokazane w komórkach `D2:F11` na rysunku 8.2.



Rysunek 8.1. Przypisanie zbioru obserwacji do zmiennej 1 daje inne wyniki niż przypisanie go do zmiennej 2

UWAGA

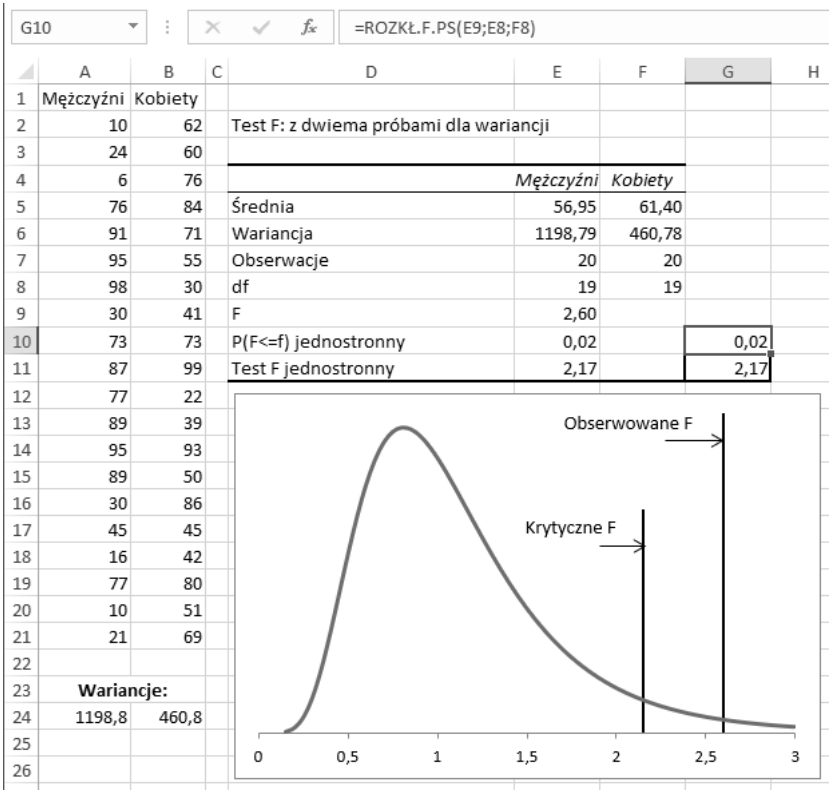
Zauważ na rysunku 8.1, że możesz zaakceptować domyślną wartość *Alfa* równą `0,05` lub zmienić ją na jakąś inną. Dokumentacja Excela, w tym dokumentacja dodatku *Analiza danych*, używa terminu „alfa" niespójnie w różnych kontekstach. W dokumentacji Excela dla narzędzia *Test F* termin ten został użyty poprawnie.

Zastosowana w narzędziu *Test F: z dwiema próbami dla wariancji* koncepcja wartości *alfa* jest taka jak opisana w rozdziale 6., „Jak zmienne są wspólnie klasyfikowane — tabele kontyngencji", oraz ponownie w rozdziale 9., „Testowanie różnic pomiędzy średnimi — podstawy". Jest to prawdopodobieństwo, że dojdziesz do wniosku, iż istnieje różnica, gdy faktycznie nie ma żadnej różnicy. W naszym przykładzie jest to prawdopodobieństwo tego, że na podstawie danych z próby dojdziesz do wniosku, iż populacje, z których je pobrałeś, mają różne wariancje, podczas gdy faktycznie są one jednakowe. To zastosowanie jest zgodne ze zwykłą statystyczną interpretacją tej wartości.

Warto także zauważyć, że dwa założenia, na których bazuje test *F* — założenie, że próby pochodzą z populacji o rozkładzie normalnym oraz że są niezależne od siebie — mają fundamentalne znaczenie. Jeżeli któreś z tych założeń nie jest spełnione, możesz założyć, że test *F* nie będzie odpowiedni.

Żadne źródło nie mówi — ani dokumentacja, ani okno dialogowe, ani inne książki na temat dodatku *Analiza danych* — że dane wyznaczone jako *Zakres zmiennej 1* w oknie dialogowym narzędzia *Test F* są zawsze traktowane jako licznik ilorazu *F*.

Rysunek 8.2. Zauważ, że wariancja zmiennej Mężczyźni jest większa niż zmiennej Kobiety w próbach losowych, a więc iloraz F jest większy niż 1,0



Narzędzie Test F zawsze dzieli zmienną 1 przez zmienną 2

Dlaczego należy to wiedzieć? Załóżmy, że postawiłeś hipotezę, iż mężczyźni mają większe od kobiet rozproszenie dowolnej cechy zmierzonej na rysunkach 8.1 i 8.2. Jeżeli uporządkowałeś dane jak na rysunku 8.2, z pomiarami dla mężczyzn w liczniku ilorazu *F*, wszystko jest w porządku. Twoja badana hipoteza jest taka, że mężczyźni mają większe rozproszenie danej cechy, a sposób skonfigurowania testu *F* odpowiada tej hipotezie. Tak skonfigurowany test pyta, czy rozproszenie cechy u mężczyzn jest na tyle większe niż u kobiet, że możesz wyeliminować przypadek — czyli błąd próbkowania — jako wyjaśnienie różnicy w ich wariancjach.

Załóżmy jednak, że nie wiedziałeś, iż narzędzie *Test F* zawsze umieszcza zmienną 1 w liczniku, a zmienną 2 w mianowniku. W takim przypadku możesz w pełni nieświadomie poinstruować narzędzie *Test F*, aby traktowało pomiary kobiet jako zmienną 1, a mężczyzn jako zmienną 2. Z danymi na rysunkach 8.1 i 8.2 otrzymałbyś wartość *F* mniejszą niż 1. Postawiłbyś hipotezę, że mężczyźni mają większe rozproszenie danej cechy, choć w rzeczywistości przetestowałbyś hipotezę przeciwną.

Tak długo, jak długo masz świadomość tego, co robisz, żadna wielka szkoda z tego nie wyniknie. Właściwa interpretacja wyników jest dość łatwa. Może jednak też prowadzić do pomyłek, szczególnie jeżeli próbujesz interpretować znaczenie wartości krytycznej raportowanej przez narzędzie *Test F*. Więcej na ten temat w następnym fragmencie.

Narzędzie Test F zmienia regułę decyzyjną

Narzędzie *Test F* zmienia sposób obliczania ilorazu *F* w zależności od tego, który zbiór danych jest identyfikowany jako zmienna 1, a który jako zmienna 2. Narzędzie zmienia także sposób przeprowadzania wnioskowania statystycznego w zależności od tego, czy obliczona statystyka *F* jest większa, czy mniejsza niż 1,0. Zwróć uwagę na wykres na rysunku 8.2. Wykres nie jest częścią wyników narzędzia *Test F*. Utworzyłem go, używając funkcji arkusza Excela `ROZKŁ.F()`.

> **WSKAZÓWKA**
>
> Jeżeli pobrałeś skoroszyty Excela do tej książki (*ftp://ftp.helion.pl/przyklady/anstae.zip*), możesz się dokładnie przyjrzeć, jak ten wykres został utworzony, otwierając skoroszyt do rozdziału 8., a w nim arkusz do rysunku 8.2.

Krzywa na wykresie reprezentuje wszystkie możliwe wartości *F*, jakie mógłbyś otrzymać, używając prób losowych po 20 obserwacji, pod warunkiem że obydwie próby pochodzą z populacji o tej samej wariancji. (Kształt rozkładu *F* zależy od liczby obserwacji w każdej próbie).

W pewnym momencie iloraz wariancji staje się tak duży, że wiara, iż obie populacje mają tę samą wariancję, staje się irracjonalna. Jeżeli populacje mają tę samą wariancję, powinieneś wierzyć, że wartość *F* różna od 1 wynika z błędu próbkowania. Nie potrzeba dużego błędu próbkowania, aby otrzymać wartość *F* równą, powiedzmy, 1,05 lub 1,10. Ale jeżeli otrzymasz próbę, której wariancja jest dwa razy taka jak innej próby — wtedy albo wystąpił nieprawdopodobnie duży stopień błędu próbkowania, albo obie populacje mają jednak różne wariancje.

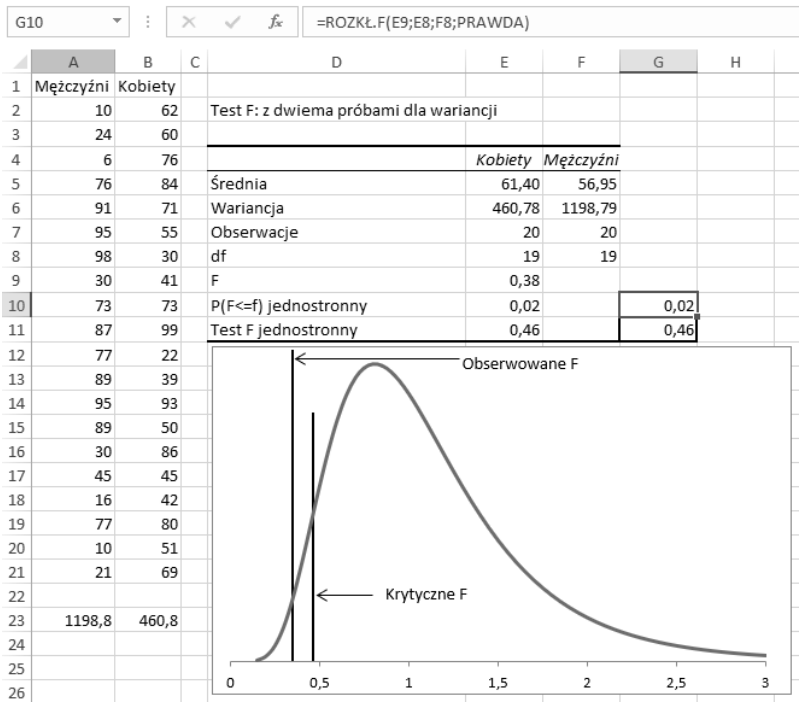
„Nieprawdopodobnie duży" jest subiektywnym określeniem. To, co jest bardzo nieprawdopodobne dla mnie, może być czymś zwykłym dla Ciebie. Dlatego badacz decyduje, co stanowi linię podziału pomiędzy nieprawdopodobnym a niewiarygodnym (często wskazaną przez koszt podjęcia błędnej decyzji). Wyrażanie tej linii podziału w terminach prawdopodobieństwa jest tu standardem. W oknie dialogowym *Test F* pokazanym na rysunku 8.1, jeżeli zaakceptujesz domyślną wartość *Alfa* równą 0,05, uznasz za niewiarygodne coś, co zdarza się co najwyżej w 5% przypadków. W przypadku testu *F* powiedziałbyś, że uznajesz za niewiarygodne otrzymanie ilorazu tak dużego, że mógłby wystąpić tylko w 5% przypadków, gdy obie populacje mają tę samą wariancję.

Wskazuje na to pionowa linia z etykietą „Krytyczne F” na rysunku 8.2. Pokazuje ona, gdzie zaczyna się największe 5% wartości *F*. Dowolna otrzymana wartość *F*, która byłaby większa niż krytyczna wartość *F*, należałaby do tych 5% i, ponieważ wybrałeś 0,05 jako kryterium *Alfa*, posłużyłaby jako dowód, że badane populacje mają inne wariancje.

Druga pionowa linia z etykietą „Obserwowane F” jest rzeczywistą wartością *F* obliczoną na podstawie danych z zakresu `A2:B21` i stanowiącą iloraz wariancji. Wariancje te są obliczone dynamicznie za pomocą funkcji `WARIANCJA.PRÓBKI()` w komórkach `A24:B24` oraz zwracane jako statyczne wartości przez narzędzie *Test F* w komórkach `E6:F6`. Narzędzie *Test F* zwraca także współczynnik *F* w komórce `E9` i jest to ta wartość, która pojawia się na wykresie jako pionowa linia z etykietą „Obserwowane F”.

Obserwowana wartość *F* równa 2,60 na rysunku 8.2 znajduje się jeszcze dalej od wartości 1,0 niż krytyczna wartość 2,17. Dlatego jeżeli użyłeś wartości α równej 0,05, Twoja reguła decyzyjna doprowadziłaby Cię do odrzucenia hipotezy, że dwie populacje podlegające próbom losowym mają równe wariancje.

Co się jednak zdarzy, jeżeli badacz, nie wiedząc, co Excel robi podczas wyznaczania wartości *F*, przypadkiem poda miary kobiet jako zmienną 1? Wtedy narzędzie *Test F* umieści wariancję kobiet 460,8 w liczniku, a wariancję mężczyzn 1198,8 w mianowniku. Wartość *F* jest teraz mniejsza niż 1,0 i otrzymasz wyniki pokazane na rysunku 8.3.



Rysunek 8.3. Teraz z mniejszą wariancją w liczniku ilorazu F wyniki są nadal znaczące, ale odwrócone

Jeżeli wiesz, co się dzieje — a wiesz — nie jest trudno stwierdzić, że obserwowana wartość *F* równa 0,38 jest tak samo nieprawdopodobna jak 2,60. Jeżeli wariancje populacji są równe, najbardziej prawdopodobnym wynikiem podzielenia wariancji jednej próby przez wariancję drugiej próby będzie wartość bliska 1,0. Patrząc na dwie wartości krytyczne na rysunkach 8.2 i 8.3, wartości 2,17 na górze i 0,46 na dole odcinają 5% obszaru pod krzywą — 5% w każdym z ogonów rozkładu. Czy umieścisz większą wariancję w liczniku, podając ją jako zmienną 1, czy też w mianowniku, podając ją jako zmienną 2, mało prawdopodobne jest uzyskanie takiego ilorazu, kiedy populacje mają równe wariancje, więc jeżeli zaakceptujesz 5% jako rozsądne kryterium, odrzucisz tę hipotezę.

Podstawy funkcji rozkładu F

Komórki `G10:G11` na rysunkach 8.2 i 8.3 zawierają funkcje arkusza, które dotyczą rozkładu *F*. Narzędzie *Test F* ich nie podaje — sam to zrobiłem — ale zauważ, że wartości pokazane w komórkach `G10:G11` są identyczne z tymi w komórkach `E10:E11`, podanymi przez narzędzie *Test F*. Jednak narzędzie *Test F* nie podaje formuł ani funkcji używanych do obliczania wyników — podaje tylko stałe wyniki. Dlatego aby lepiej zrozumieć, jak działa dane narzędzie, na przykład *Test F* z dodatku *Analiza danych*, warto znać i rozumieć używane przez nie funkcje arkusza.

Komórka `G10` na rysunku 8.2 zawiera formułę:

```
=ROZKŁ.F.PS(E9;E8;F8)
```

Funkcja `ROZKŁ.F.PS` zwraca prawdopodobieństwo, które możesz interpretować jako pole powierzchni pod krzywą. Sufiks `PS` w nazwie funkcji mówi Excelowi, że potrzebne jest pole powierzchni pod prawym ogonem krzywej. Gdybyś użył zamiast tego funkcji `ROZKŁ.F()`, Excel zwróciłby pole powierzchni pod lewym ogonem krzywej.

Pierwszym argumentem funkcji, tutaj `E9`, jest wartość *F*. Używana jako argument funkcji `ROZKŁ.F.PS()` wartość w komórce `E9` powoduje wyliczenie pola powierzchni obszaru pod krzywą, który leży po prawej stronie tej wartości. Na rysunku 8.2 wartość w `E9` to 2,60, więc Excel zwraca 0,02 — 2% obszaru pod krzywą leży po prawej stronie wartości *F* równej 2,60.

Zgodnie z opisem w poprzednim fragmencie kształt rozkładu *F* zależy od liczby obserwacji, które kształtują wariancję w liczniku i mianowniku ilorazu *F*. Ściśle rzecz biorąc, używasz tutaj stopni swobody zamiast rzeczywistej liczby obserwacji — stopnie swobody to liczba obserwacji minus 1. Drugi i trzeci argument funkcji `ROZKŁ.F.PS()` to odpowiednio stopnie swobody w liczniku i mianowniku.

Na podstawie wyników zwracanych przez tę funkcję możesz dojść do wniosku, że zakładając taką samą wariancję populacji mężczyzn i kobiet, stwierdziłbyś wartość *F* równą co najmniej 2,60 w tylko 2% próbek, które mogły zostać wzięte z tych populacji. Możesz uznać za bardziej rozsądny wniosek, że założenie o równych wariancjach populacji jest nieprawidłowe, niż stwierdzenie, że otrzymałeś dosyć nieprawdopodobną wartość *F*.

Formuła w komórce G11 na rysunku 8.2 ma postać:

```
=ROZKŁ.F.ODWR(0,95;E8;F8)
```

Zamiast zwracać pole powierzchni pod krzywą, jak funkcje ROZKŁ.F() i ROZKŁ.F.PS(), funkcja ROZKŁ.F.ODWR() przyjmuje pole powierzchni jako argument i zwraca wartość *F*. Tutaj drugi i trzeci argument w komórkach E8 i F8 są takie same jak w przypadku funkcji ROZKŁ.F.PS() — liczba stopni swobody w liczniku i mianowniku. Argument 0,95 instruuje Excela, że potrzebna jest wartość *F* odpowiadająca 95% pola powierzchni pod krzywą. Funkcja zwraca 2,17 w komórce G11, więc 95% krzywej leży po lewej stronie wartości 2,17 w rozkładzie *F* z 19 stopniami swobody w liczniku i mianowniku. Narzędzie *Test F* zwraca tę samą wartość jako wartość w komórce E11.

(Sufiks ODWR funkcji jest skrótem od słowa *odwrotny*. Funkcja, która zwraca wartość statystyki jest zgodnie z konwencją uważana za funkcję odwrotną do funkcji mierzącej pole powierzchni pod krzywą).

Porównaj funkcje na rysunku 8.2, które zostały już opisane, z wersjami na rysunku 8.3. Tutaj w komórce G10 znajduje się formuła:

```
=ROZKŁ.F(E9;E8;F8;PRAWDA)
```

Tym razem zamiast funkcji ROZKŁ.F.PS() została użyta funkcja ROZKŁ.F(). Funkcja ROZKŁ.F() zwraca pole powierzchni obszaru pod krzywą po *lewej* stronie wartości *F*, którą podałeś (tutaj wartość wynosi 0,38, co jest wartością komórki E9, ilorazu wariancji kobiet przez wariancję mężczyzn).

> **UWAGA**
>
> Funkcja ROZKŁ.F() przyjmuje czwarty argument, którego nie przyjmuje funkcja ROZKŁ.F.PS(). W funkcji ROZKŁ.F() możesz podać wartość PRAWDA, jak wyżej, aby zażądać pola powierzchni po lewej stronie wartości *F*. Jeżeli zamiast tego podasz FAŁSZ, Excel zwróci wysokość krzywej w punkcie *F*. Poza innymi zastosowaniami wartość wysokości jest nieodzowna do przedstawienia rozkładu *F* na wykresie. Podobne rozważania dotyczą kreślenia rozkładów normalnych, rozkładów t-Studenta, rozkładów χ^2 itd.

Porównując wykresy na rysunkach 8.2 i 8.3, możesz się łatwo przekonać, że otrzymanie ilorazu 0,38 (wariancji kobiet do mężczyzn) i otrzymanie ilorazu 2,60 (wariancji mężczyzn do kobiet) jest tak samo nieprawdopodobne. Jednak może Cię zmylić fakt, że w tych dwóch zbiorach wyników krytyczna wartość jest różna. Na rysunku 8.2 jest to 2,17, ponieważ narzędzie *Test F* działa z wartością *F* większą niż 1,0, więc pytanie dotyczy tego, o ile większa od 1,0 musi być obserwowana wartość *F*, aby odciąć górne 5% rozkładu (lub inną wybraną wartość α zamiast 0,05).

Na rysunku 8.3 krytyczna wartość wynosi 0,46, ponieważ narzędzie *Test F* działa z wartością *F* mniejszą niż 1,0, więc pytanie dotyczy tego, o ile mniejsza od 1,0 musi być obserwowana

wartość *F*, abyś uznał ją za nieprawdopodobnie małą — mniejszą niż najmniejsze 5% obserwowanych wartości, jeżeli populacje mają tę samą wariancję.

Ta krytyczna wartość równa 0,46 w komórce `G11` rysunku 8.3 jest zwracana przez następującą formułę:

```
=ROZKŁ.F.ODWR(0,05;E8;F8)
```

Podczas gdy, jak wcześniej wspomniałem, formuła w komórce `G11` rysunku 8.2 ma postać:

```
=ROZKŁ.F.ODWR(0,95;E8;F8)
```

W tej ostatniej wersji funkcja zwraca wartość *F*, która odcina dolne 95% obszaru pod krzywą — dlatego większe wartości mają co najwyżej 5% szans wystąpienia.

W pierwszej wersji funkcja zwraca wartość *F*, która odcina dolne 5% obszaru pod krzywą. Jest to wartość krytyczna, jeżeli tak ustanowiłeś obserwowaną wartość *F*, by mniejsza wartość była w liczniku.

Istnieje funkcja `ROZKŁ.F.ODWR.PS()`, której możesz użyć zamiast `=ROZKŁ.F.ODWR(0,95;E8;F8)`. Jest to po prostu sprawa osobistych upodobań. Funkcja `ROZKŁ.F.ODWR.PS()` zwraca wartość *F*, która odcina prawy ogon, a nie lewy, jak w przypadku funkcji `ROZKŁ.F.ODWR()`. Dlatego poniższe dwie funkcje są równoważne:

```
=ROZKŁ.F.ODWR(0,95;E8;F8)
```

i

```
=ROZKŁ.F.ODWR.PS(0,05;E8;F8)
```

> **UWAGA**
>
> Narzędzie *Test F* **nie** pokazuje wykresu. Dobrym pomysłem jest wyświetlenie wyników testu na wykresie, abyś lepiej zrozumiał, co się dzieje, ale musisz go skonstruować samodzielnie. Pobierz skoroszyt z witryny wydawcy, aby zobaczyć, jak zdefiniować wykres.

Stawianie hipotezy dwustronnej

Do tej pory interpretowaliśmy wyniki narzędzia *Test F* pod względem dwóch wzajemnie wykluczających się hipotez:

- Pomiędzy dwiema populacjami nie ma różnicy mierzonej ich wariancjami.
- Populacja mężczyzn ma większą wariancję niż populacja kobiet.

Druga hipoteza jest nazywana hipotezą **kierunkową**, ponieważ określa, że któraś z dwóch wariancji jest zgodnie z oczekiwaniami większa. (Jest ona także nazywana, cokolwiek niedbale, hipotezą **jednostronną**, ponieważ przywiązujesz wagę tylko do jednego ogona rozkładu. Jest to dość niedbałe określenie, ponieważ — jak zobaczysz w dalszych rozdziałach — wiele dwustronnych hipotez odnosi się tylko do jednego ogona w rozkładzie *F*).

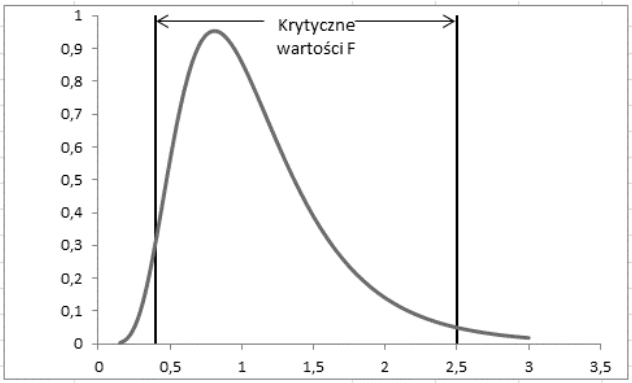
A co, jeżeli nie chcesz z góry określać, która wariancja jest większa? Wtedy Twoje dwie wzajemnie się wykluczające hipotezy (w podręcznikach zwane **zespołem hipotez**) mogą być następujące:

- *Nie ma żadnej* różnicy pomiędzy dwiema populacjami pod względem zmierzonych wariancji.
- *Istnieje* różnica pomiędzy dwiema populacjami pod względem zmierzonych wariancji.

Zauważ, że druga hipoteza nie zakłada, która populacja ma większą wariancję — a po prostu, że wariancje dwóch populacji są różne. Jest to hipoteza **dwustronna**. Ma to duży wpływ na sposób parametryzacji i interpretacji testu *F* (i testów *t*, jak zobaczysz w rozdziałach 10. i 11.).

Podejście graficzne

Rysunek 8.4 pokazuje, w jaki sposób testy dwustronne różnią się od testów jednostronnych pokazanych na rysunkach 8.2 i 8.3.



Rysunek 8.4. W przypadku hipotezy dwustronnej obszar α jest podzielony pomiędzy dwa ogony

W przypadku podobnym do pokazanego na rysunku 8.4 nie przyjmujesz z góry, która populacja ma większą wariancję, a tylko to, że są one nierówne. Dlatego jeżeli zdecydujesz, że jesteś gotów uznać wyniki z 5-procentowym prawdopodobieństwem za wystarczająco nieprawdopodobne, aby odrzucić hipotezę zerową, wtedy 5% prawdopodobieństwa musi być dzielone pomiędzy oba ogony rozkładu. Dolny i górny ogon dostają po 2,5%. (Oczywiście możesz zdecydować, że do odrzucenia hipotezy niezbędne jest nie 5%, a 1% lub dowolna inna wartość, którą Twój osobisty i zawodowy osąd uzna za „nieprawdopodobną". Zwróć jednak uwagę, że w teście dwustronnym dzielisz ten nieprawdopodobny poziom α pomiędzy dwa ogony rozkładu).

Jedną z konsekwencji przyjęcia dwustronnej hipotezy alternatywnej jest to, że wartości krytyczne są bliżej ogonów rozkładu. Na rysunku 8.4 dwustronna hipoteza przesuwa górną wartość krytyczną do około 2,5, podczas gdy na rysunku 8.2 hipoteza kierunkowa umieściła wartość krytyczną na poziomie 2,17. (To tylko przypadek, że górna wartość krytyczna wynosi około 2,5 i odcina 2,5% pola powierzchni).

Powodem przesunięcia się wartości krytycznej jest to, że na rysunku 8.4 wartości krytyczne odcinają dolne i górne 2,5% rozkładów zamiast dolnego 5% lub górnego 5%, jak na rysunkach 8.2 i 8.3. Dlatego wartości krytyczne są bardziej oddalone od środka rozkładu na rysunku 8.4.

Stosowanie narzędzia Test F do hipotezy dwustronnej

Jeżeli chcesz użyć dwustronnej hipotezy, zmniejsz poziom α o połowę. Dostosuj go w oknie dialogowym narzędzia *Test F*. Jeżeli chcesz, aby ogólny poziom α wynosił 5%, wpisz **`0,025`** w oknie tego narzędzia.

Określenie poziomu α wpływa **tylko** na krytyczną wartość F zwracaną przez narzędzie *Test F*. Możesz zawsze spojrzeć na wartość p obserwowanej wartości F zwracaną przez narzędzie (na przykład w komórce `E10` na rysunku 8.3), która nie ulegnie zmianie. Następnie zdecyduj, czy wartość p jest wystarczająco niska, aby uznać za nieprawdopodobną hipotezę, że ten wynik jest spowodowany błędem próbkowania. W praktyce zależy to od tego, czy chcesz myśleć w kategoriach prawdopodobieństw (przywiązać uwagę do wartości α i p), czy w kategoriach wartości F (zastanawiać się nad obserwowaną i krytyczną wartością F).

Pułapka do ominięcia

Tym, czego *nie* wolno zrobić, jeżeli postawiłeś *hipotezę dwustronną*, jest sugerowanie się danymi przy decyzji o tym, którą grupę danych umieścić w liczniku ilorazu F, używając tej grupy jako zmiennej 1 w oknie dialogowym narzędzia *Test F*.

Możesz zdecydować — przed zerknięciem w dane — że będziesz traktować dowolną grupę o większej wariancji jako zmienną 1. Nie: „Widzę, że mężczyźni mają większą wariancję, więc potraktuję ich dane jako zmienną 1”, ale: „Umieszczę dowolną grupę o większej wariancji w liczniku ilorazu F, wyznaczając tę grupę jako zmienną 1”.

Nie ma też nic złego w tym, aby przypisać jeden z dwóch zbiorów danych do zmiennej 1 za pomocą rzutu monetą lub innego źródła losowości.

Jeżeli zdecydujesz, że zawsze umieścisz większą wariancję w liczniku ilorazu F, nigdy nie otrzymasz wartości F mniejszej niż 1,0. Pytasz jednak zarówno o górny ogon rozkładu, jak i o dolny. Dlatego jeżeli test jest dwustronny, na pewno musisz umieścić w oknie dialogowym połowę wartości α, którą chcesz uwzględnić. Zauważ, że jest to spójne z radą, jaką dałem we wcześniejszym fragmencie, aby podać połowę wartości α, którą rzeczywiście chcesz otrzymać, gdy masz do czynienia z oknem dialogowym narzędzia *Test F* i stawiasz dwustronną hipotezę alternatywną.

Dostępne wybory

Podsumowując, sposób określenia parametrów w narzędziu *Test F* dodatku *Analiza danych* zależy od tego, czy stawiasz hipotezę jednostronną, czy dwustronną. W następnych dwóch fragmentach krótko opiszę każdą alternatywę przy założeniu, że podałeś α — prawdopodobieństwo, iż obserwowane wyniki są przypadkowe — równe 0,05.

Hipoteza kierunkowa

Postaw hipotezę kierunkową, jeżeli znajduje ona pokrycie w jakiejkolwiek teorii. Jeżeli teoria głosi, że mężczyźni powinni mieć większą wariancję pewnej cechy niż kobiety, niech hipoteza alternatywna będzie kierunkowa: mężczyźni mają większą wariancję danej cechy. Użyj okna dialogowego narzędzia *Test F* do umieszczenia wariancji mężczyzn w liczniku ilorazu *F* (przypisz dane mężczyzn zmiennej 1) i podaj wartość *Alfa* równą 0,05. Uznaj swoją hipotezę alternatywną za prawidłową tylko wtedy, gdy obserwowana wartość *F* przekracza krytyczną wartość *F*.

Nie odrzucaj hipotezy zerowej o braku różnicy, nawet jeżeli wariancja próby losowej mężczyzn jest znacznie mniejsza niż kobiet. Gdy raz postawisz hipotezę kierunkową, która wskazuje na konkretny kierunek, musisz się jej trzymać. Postawienie hipotezy kierunkowej po analizie danych byłoby zwykłym nagięciem faktów do wniosków.

Jeżeli jednak będziesz uważnie stosował się do zasad rządzących formułowaniem hipotez kierunkowych, zwiększysz moc testu *F* służącego do ich weryfikacji (w porównaniu z mocą tego testu dla hipotez dwustronnych), a więc częściej prawidłowo potwierdzisz, że różnica między średnimi w populacji faktycznie istnieje.

Hipoteza dwustronna

Postaw hipotezę dwustronną, że próbkowane populacje mają różne wariancje, ale nie określaj, która jest większa. Dla wygody traktuj grupę z większą wariancją jako zmienną 1, przedziel wartość α na pół, gdy będziesz wypełniać pola okna dialogowego, i uruchom test *F* jednokrotnie. Jeżeli wyliczona wartość *p* jest mniejsza niż nominalna wartość α, przyjmij hipotezę alternatywną, że populacje mają różne wariancje.

Zignoruj etykietę wyników testu *F* „P(F<=f) jednostronny". Ta etykieta jest sama w sobie myląca, symbole są niezdefiniowane i pozostaje niezmieniona bez względu na to, czy otrzymana wartość *F* jest większa, czy mniejsza od wartości krytycznej. Co więcej, prawdopodobieństwo tego, że jedna liczba jest mniejsza bądź równa drugiej wynosi albo 1, albo 0. Bo albo ta nierówność zachodzi, albo nie. Wartości *F* oraz *f* są takimi właśnie dwiema liczbami, dlatego też stwierdzenia w rodzaju „Prawdopodobieństwo, że 2,6 jest większe niż 2,17, wynosi 0,02" są pozbawione sensu.

I na odwrót: w wynikach narzędzia *Test F* wielkość z etykietą „P(F<=f) jednostronny" jest prawdopodobieństwem uzyskania obserwowanej wartości *F* przy założeniu, że populacje, z których zostały pobrane próby losowe, mają taką samą wariancję.

Replikowalność badań

Od mniej więcej 2015 roku w prasie branżowej zaczęto zwracać uwagę na pewien zaskakujący fakt, utrudniający interpretację i ocenę wyników badań empirycznych (wątpliwości te przedostały się następnie do prasy popularnej). Chodzi o kwestię **replikowalności** (powtarzalności) wyników badań. The Center for Open Science z siedzibą w amerykańskim stanie Wirginia włożyło sporo wysiłku w próbę powtórzenia wyników uzyskanych w trakcie badań empirycznych przeprowadzonych w przeszłości, które uznano za kluczowe dla rozwoju takich dziedzin jak psychologia czy medycyna (w tym drugim przypadku chodziło głównie o badania nad przerzutami nowotworów oraz najbardziej skutecznymi sposobami ich zwalczania).

Do czasu gdy piszę te słowa, uczestnicy projektu nie byli w stanie opublikować żadnych jednoznacznych wniosków ze swoich badań. Ogłosili mianowicie, że jak dotąd nie udało im się zreplikować wyników uzyskanych w licznych badaniach z zakresu psychologii i medycyny, z drugiej jednak strony przyznali, że byli w stanie potwierdzić wnioski uzyskane w ramach innych badań.

Nie wdając się w szczegóły, jest to efekt dość nieoczekiwany. Okazuje się, że w badaniach, z których wnioski udało się powtórzyć, wnioskowanie prowadzone było na tzw. poziomie istotności z przedziału od 0,05 do 0,001. Jak wyjaśnię to w dalszej części książki, informacja ta oznacza, że udało się zgromadzić wystarczające dowody, by uznać, że uzyskane rezultaty w zaledwie 5 procentach (lub odpowiednio: 1 procencie czy 0,01 procenta) przeprowadzonych w porównywalnych warunkach badań można by uznać za przypadkowe.

Tymczasem uczestnicy projektu stwierdzili, że w badaniach z zakresu psychologii jedynie w 47 procentach przypadków (około połowy) wielkość efektu (ang. *effect size*) mieściła się w 95-procentowym przedziale ufności zbudowanym wokół wyników, które udało się zreplikować (zgodnie z definicją przedziału ufności należałoby oczekiwać, że wielkość efektu będzie się mieściła w granicach takiego przedziału w ok. 95 procentach przypadków, a nie 47 procentach; wielkość efektu to standaryzowana miara wyników badań empirycznych, którą można zdefiniować np. jako stosunek różnicy (dystansu) między średnimi wartościami badanej cechy w dwóch grupach do odchylenia standardowego w próbie czy choćby zwykłego współczynnika korelacji).

Pozostało przy tym zagadką, dlaczego nie udało się zreplikować wyników dość znaczącego odsetka badań z zakresu medycyny, choć z drugiej strony porównywalny odsetek innych badań znalazł potwierdzenie we wnioskach płynących z powtórzonych badań.

Taka sytuacja może wręcz rodzić frustracje, gdy okazuje się, że nie udaje się potwierdzić wniosków (czy wręcz przeczy się im), które uzyskano w wyniku dobrze przemyślanego (przynajmniej w teorii) badania. Replikowalność jest jednak bardzo ważnym narzędziem tzw. metody naukowej, które chroni nas przed zbyt pochopnym akceptowaniem wyników zgodnych ze stanem naszej wiedzy.

Zanim powstanie końcowy raport podsumowujący kompleksowo wysiłki zespołu z Wirginii, minie trochę czasu. Już dziś jednak na podstawie cząstkowych wyników tego eksperymentu możemy sformułować hipotezę, że przyczyną uniemożliwiającą pełną replikację wyników badań dokonanych w przeszłości może być problem odtworzenia warunków, w jakich te badania były prowadzone. Badania z zakresu biochemii są niezwykle skomplikowane, przez co łatwo jest nieumyślnie naruszyć założenia leżące u ich podstaw. To zaś sprawia, że wtórne badanie nie będzie dokładnym powtórzeniem badania pierwotnego, nawet jeżeli tak by się nam wydawało. Czy możemy więc wyciągnąć jakieś wnioski z tego eksperymentu, zanim doczekamy się finalnego raportu?

W części tego rozdziału poświęconej wewnętrznej trafności badań stwierdziliśmy, że w wielu z nich porównuje się średnią wartość analizowanej cechy w grupie eksperymentalnej z jej odpowiednikiem w próbie kontrolnej (średnie te są zwykle liczone z wyników pomiaru odpowiedniego dla stawianej hipotezy). Choć jest to prawda, chciałbym teraz nawiązać do materiału, który znajdziesz w dalszych rozdziałach tej książki, uprzedzając nieco fakty: jak się przekonasz, jest możliwa analiza efektów dwóch lub większej liczby czynników w ramach jednego eksperymentu.

Na przykład możemy łatwo rozszerzyć zakres badania, którego celem jest zbadanie wielkości efektu w zakresie skuteczności pewnego leku (względem efektu działania placebo) poprzez dodanie kolejnego czynnika, jakim może być płeć pacjenta. W takim przypadku musielibyśmy porównać więcej niż tylko dwie średnie, zestawiając ze sobą: grupę eksperymentalną z grupą kontrolną oraz kobiety z mężczyznami, a także średnie w każdej z czterech grup: kobiety w grupie eksperymentalnej, kobiety w grupie kontrolnej, mężczyźni w grupie eksperymentalnej, mężczyźni w grupie kontrolnej. Wartość takiego eksperymentu będzie wyższa od wartości dwóch niezależnych badań, z których każde analizowałoby pojedynczą cechę. I to z wielu powodów, które omawiam szczegółowo w następnych rozdziałach. W tym miejscu skupię się na jednym z nich: dodaniu nowego *rodzaju* zmiennej.

Wyobraź sobie, że prowadzisz eksperyment, który ma za zadanie zbadać wielkość efektu związanego z podaniem nowego leku losowo wybranej grupie dorosłych pacjentów w porównaniu z podaniem placebo innej grupie losowo wybranych dorosłych pacjentów. Badacze określają taką zmienną (nowy lek kontra placebo) jako **czynnik stały**. Twoim celem jest bowiem wyłącznie porównanie efektów zastosowania konkretnego leku ze skutkami zażycia placebo. Nie traktujesz tego leku jako jednoelementowej próbki z populacji wielu możliwych leków. Twoja uwaga skupia się **stale** na tym leku i tylko na nim.

Wyobraź sobie teraz, że możesz przeprowadzić testy takiego leku w kilku niezależnych laboratoriach. Eksperyment prowadzony przez ekspertów z The Center for Open Science uzmysłowił Ci, że niewielkie różnice w warunkach prowadzenia eksperymentu mogą skutkować całkowicie odmiennymi wnioskami z badania. Zdecydowałeś więc, że powtórzysz badanie w dziesięciu różnych laboratoriach.

Rozumiesz zatem, że możesz uzyskać różne wyniki z każdego z tych laboratoriów. Mogą one być nieistotne statystycznie, ale też mogą znacząco wpłynąć na końcowe wnioski. Co więcej, nie będziesz tak bardzo zaskoczony, gdy okaże się, że kolejne z badań nie są w stanie zreplikować wyników poprzednich.

W tak zdefiniowanym podejściu dwuczynnikowym — efekt działania leku w zależności od laboratorium — Twoja uwaga jest nadal skupiona na badaniu efektów zażycia konkretnego leku w porównaniu z efektem placebo. Ale z założenia (z góry) chciałbyś dodatkowo uogólnić wnioski z takiego badania na dowolne laboratorium (eliminując wpływ czynnika miejscowego). W tym celu traktowałbyś wybrane 10 z nich jako losową próbę pobraną z populacji wszystkich laboratoriów biochemicznych świata. Badacze mówią w tym kontekście o **czynniku losowym**. To, czy daną zmienną traktujesz jako czynnik stały, czy losowy, wpływa bezpośrednio na przebieg badania i sposób interpretacji wniosków. (Więcej miejsca czynnikom stałym i losowym poświęcam w rozdziale 13., „Planowanie eksperymentu a ANOVA").

W pierwszych latach XX wieku, kiedy używane przez nas po dziś dzień metody analizy statystycznej dopiero powstawały, stosowano je w praktyce głównie w rolnictwie. Badano skuteczność pewnych specyfików czy technik — nawozów, środków owadobójczych, rozplanowania kanałów irygacyjnych itp. — na masową skalę, zatem i uzyskiwane wyniki mierzone były w dużych jednostkach, np. plony liczone w setkach kwintali. Nawet jednak przy tych uwarunkowaniach największe umysły epoki potrafiły zaproponować techniki, które pozwalały uwzględniać efekty tzw. zmiennych uciążliwych.

Badanie wpływu zmiennych uciążliwych — na przykład różnic między laboratoriami — ma sens także i dziś, gdy w ramach prowadzonych badań analizujemy wyniki tak precyzyjnych pomiarów, jak skala penetracji komórek rakowych przez peptydy (w celu zwiększenia efektywności innych środków mających zniszczyć komórki rakowe). Możesz sobie wyobrazić, jak łatwo jest w takiej sytuacji dojść do wniosku o niereplikowalności wyników pewnych badań, gdy znaczenie może mieć *n*-ta cyfra po przecinku.

Jak wspomniałem powyżej, dodanie czynnika losowego, takiego jak przypadkowo wybrane laboratorium, zmienia naturę samego eksperymentu i charakter prowadzonych analiz statystycznych. Nadal jednak nie jest to zmiana, którą można by porównać w skutkach do fundamentalnej zmiany w planie eksperymentu. Te dwa obszary, analiza statystyczna i planowanie eksperymentów, są ze sobą ściśle powiązane. Zwracam na to uwagę w rozdziale 13. Chcę Cię jednocześnie zachęcić, byś na zawsze zapamiętał, że nawet niewielka zmiana w planie eksperymentu może mieć daleko idące konsekwencje dla przeprowadzanych analiz i wniosków z nich płynących.

Uwagi końcowe

W tym rozdziale starałem się zwrócić Twoją uwagę na to, jak różnorodne są przypadki, gdy analiza statystyczna może zwieść Cię na manowce, i w jaki sposób możesz się przed tym obronić. Co więcej, metoda naukowa nakazuje nam kontrolować możliwe przyczyny uzyskania takich, a nie innych wyników eksperymentu, mogących nierzadko wykraczać poza czynniki, które były dla nas faktyczną motywacją do jego przeprowadzenia. Z punktu widzenia planowania eksperymentów faktyczna analiza statystyczna i elementy jej kontroli są w gruncie rzeczy nieistotne. Kontrolowanie wpływu przypadkowości na wnioski z badań jest oczywiście sprawą ważną, niemniej przypadkowość jest tylko jednym z wielu zagrożeń, jakie mogą podważyć trafność (ważność) eksperymentu.

Nie wolno Ci również zapominać, jak ważne jest szczegółowe rozpoznanie narzędzi, które w ogóle pozwalają Ci przeprowadzić jakąkolwiek analizę statystyczną. W drugiej części rozdziału starałem się uczulić Cię na to, jak bardzo musisz mieć się na baczności, by uniknąć pułapek zastawionych przez autorów aplikacji statystycznych, z których korzystasz — bez względu na to, na którą się zdecydowałeś. Dokumentacja oprogramowania do analiz statystycznych bywa uboga, szczególnie gdy ograniczysz się do jej wersji oficjalnej, dostarczanej przez producenta. Wiele potencjalnych pułapek w ogóle jest w niej pomijanych. Twoim najlepszym orężem w tej nierównej walce będzie pozyskanie gruntownej wiedzy z zakresu teorii statystyki uzupełnione o praktyczne eksperymenty z oprogramowaniem — by oswoić się z wszystkimi jego dziwactwami.

Tyle tytułem podsumowania. W kolejnym rozdziale dowiesz się, w jaki sposób wnioskowanie statystyczne w ogólności, a z użyciem Excela w szczególności, pozwala nam potwierdzać istnienie różnic między średnimi w grupach lub mu zaprzeczać.

Skorowidz

D

E

F

G

H

I

J

K

L

M

N

O

P

R

S

Ś

T

U

W

Z

ANALIZA STATYSTYCZNA: Z EXCELEM TO ŁATWE!

Microsoft Excel jest naprawdę wszechstronnym narzędziem. Umożliwia tworzenie raportów, inteligentnych modeli, a także prowadzenie złożonych analiz statystycznych. Profesjonalni badacze, studenci i biznesmeni, którzy zajmują się analizą danych i statystyką, właśnie Excela traktują jako ulubione narzędzie pracy. Osiągnięcie biegłości w posługiwaniu się nim wymaga odrobiny wysiłku, jednak zdobyta w ten sposób wiedza okazuje się bardzo przydatna!

Niniejsza książka jest praktycznym przewodnikiem po analizie statystycznej i funkcjach statystycznych Excela. Dzięki licznym przykładom nauczysz się dobierać właściwe narzędzia do rozwiązania konkretnego problemu. Dowiesz się, jak korzystać z korelacji, regresji oraz analizy wariancji i kowariancji. Zastosujesz Excela do testowania hipotez statystycznych z zastosowaniem rozkładów normalnych, dwumianowych, t-Studenta i F-Snedecora. Zapoznasz się ze znaczeniem najważniejszych pojęć statystycznych i unikniesz typowych błędów. W tym wydaniu książki uwzględniono nowe funkcje, które pojawiły się w Excelu 2016. Dowiesz się, do czego służą i jak je można wykorzystać.

Najważniejsze zagadnienia ujęte w książce:

- podstawowe pojęcia w statystyce, statystyce opisowej i wnioskowaniu statystycznym
- ważniejsze funkcje statystyczne Excela
- testy *z*, testy *t* i dodatek *Analiza danych Excela*
- histogramy i wykresy w Excelu
- analiza wariancji i kowariancji oraz metody regresji

Conrad Carlberg jest uznanym autorytetem w dziedzinie statystyki i analizy danych. Doskonale zna takie aplikacje jak MS Excel, SAS i Oracle. Wielokrotnie otrzymywał nagrodę MVP. Od ponad 20 lat doradza firmom, które chcą podejmować decyzje biznesowe na podstawie analizy danych. Z radością pisze o tych technikach, ze szczególnym upodobaniem dzieląc się swoją ogromną wiedzą o MS Excelu, którego uważa za najpopularniejszy w świecie program do analiz numerycznych.

Helion
helion.pl
HELION SA
ul. Kościuszki 1c
44-100 Gliwice
tel.: 32 230 98 63
helion@helion.pl

KOD KORZYŚCI
Sięgnij po więcej!

ISBN 978-83-283-4467-9

9 788328 344679

Pearson

Cena: 89,00 zł