Psychologia Społeczna
2016 tom 11 4 (39) strony 458–473
ISSN 1896-1800
doi: 10.7366/1896180020163906

W statystycznym matriksie: kontrowersje wokół testowania istotności hipotezy zerowej (*null hypothesis significance testing*, NHST) oraz p-wartości

Lilianna Jarmakowska-Kostrzanowska
Interdyscyplinarne Centrum Nowoczesnych Technologii, Uniwersytet Mikołaja Kopernika

Wydaje się, że analiza danych w wielu badaniach psychologicznych opiera się w dużej mierze na testowaniu istotności hipotezy zerowej (NHST), czyli kilkustopniowej procedury, która jest niefortunnym połączeniem dwóch niezgodnych ze sobą podejść – Fishera oraz Neymana–Pearsona. Mimo nawoływania do jej porzucenia, badacze, w tym psychologowie, nadal posługują się tą wadliwą procedurą. Niniejszy artykuł przedstawia koncepcje, z jakich czerpie NHST, oraz kontrowersje wokół jej najbardziej znanego elementu, p-wartości. Tekst porusza zarówno zarzuty stawiane tej statystyce (np. zależność od liczebności), jak i wyjaśnia jej interpretację (np. jako prawdopodobieństwo warunkowe). Celem artykułu jest w sposób przystępny dla polskich badaczy z obszaru nauk społecznych przedstawiać zarys problematyki NHST.

Słowa kluczowe: *testowanie istotności hipotezy zerowej (NHST)*, *p-wartość*, *testowanie hipotez*

Testowanie istotności hipotezy zerowej (*Null Hypothesis Significance Testing* [*Procedure*]; NHST lub NHSTP) jest procedurą zarówno szeroko rozpowszechnioną w świecie badawczym, jak i bardzo krytykowaną (np.: Cohen, 1994). Uznawana jest za połączenie dwóch nieprzystających do siebie podejść i wciąż nawołuje się do jej porzucenia. Mimo tych zaleceń, zeszłoroczny, lutowy artykuł wstępny *Basic and Applied Social Psychology* (BASP) wzbudził wiele kontrowersji, wprost *zakazując* NHSTP (Trafimow, Marks, 2015). Zgodnie z wytycznymi artykuły ukazujące się w tym periodyku w kolejnych numerach będą pozbawione jakichkolwiek oznak NHST, przede wszystkim p-wartości, *t*-wartości, *F*-wartości oraz przedziałów ufności. Decyzja ta wywołała poruszenie, nawet wśród osób niezwiązanych z psychologią. W celu oceny, czy polityka redakcyjna BASP jest właściwa i czy inne wydawnictwa, a przede wszystkim badacze, w tym psychologowie, powinni pójść tą drogą i zrezygnować z testowania istotności hipotezy zerowej (NHST) oraz wszelkich jej elementów pochodnych (takich jak p-wartość), należy najpierw poznać, czym ona sama jest. Celem niniejszego artykułu jest zatem przedstawienie badaczom z dziedziny nauk społecznych w sposób przystępny zarysu problematyki NHST.

Artykuł jest zorganizowany w następujący sposób: pierwsza sekcja prezentuje historyczno-filozoficzne tło koncepcji, sylwetki ich autorów, a także najważniejsze punkty podejść; w sekcji drugiej zaprezentowane są błędne przekonania o p-wartości, które nadal funkcjonują w środowisku naukowym; sekcja trzecia zawiera zarzuty stawiane p-wartości; sekcja czwarta stanowi podsumowanie.

Kontekst historyczny

W przeważającej większości przypadków narzędzia statystyczne można poznawać w całkowitym oderwaniu od okoliczności, w jakich powstały. Procedura NHST nie stanowi wyjątku od tej reguły – od kilkudziesięciu lat jest z powodzeniem wykorzystywana jako algorytm

Lilianna Jarmakowska-Kostrzanowska, Interdyscyplinarne Centrum Nowoczesnych Technologii, Uniwersytet Mikołaja Kopernika, ul. Wileńska 4, 87-100 Toruń,
e-mail: lilianna.kostrzanowska@gmail.com
Pomoc, której nie sposób przecenić, udzieliły mi następujące osoby (w kolejności alfabetycznej): mgr Anna Bańbura, dr Tomasz Jarmakowski-Kostrzanowski, mgr Michał Nowak, dr hab. Anna Paszkowska-Rogacz, Michał Wasążnik. Bardzo Wam dziękuję za poświęcony czas!

postępowania przez kolejne pokolenia psychologów, którzy na ogół nie uświadamiają sobie jej proweniencji. Trzeba jednak podkreślić, że znajomość kontekstu historycznego lub filozoficznego pozwoliłaby zarówno przestać traktować NHST w kategoriach czarnej skrzynki, jak i ułatwiłaby zrozumienie powodów tak szerokiej krytyki.

Niezgodność czy też niekompatybilność podejść wchodzących w skład NHST ma źródło w rozbieżnych rozumieniach pojęcia prawdopodobieństwa Fishera oraz Neymana. Próba odpowiedzi na pytanie, dlaczego koncepcje, z jakich czerpie NHST, są ze sobą niekompatybilne, napotyka dwa rodzaje trudności. Po pierwsze odwołuje się do wiedzy filozoficznej, gdyż prace matematyczne pisane przed stu laty były bardzo mocno osadzone w filozoficznych rozważaniach. Mimo że – jak pisał Neyman (1933) – zagadnienie testowania hipotez jest starym problemem, najpierw należy ustalić sposób interpretacji pojęcia *prawdopodobieństwa*, który jest bazą dla statystyki. Jak przyznają Halpin i Stam (2006), w historii matematyki ciągle toczy się debata nad interpretacją pojęcia prawdopodobieństwa, np. frekwentyści opierają się na częstościowej definicji prawdopodobieństwa von Misesa (Dienes, 2008), gdzie prawdopodobieństwo zdarzenia A to graniczna częstość jego występowania. Zwolennicy Bayesa przypisują prawdopodobieństwa według subiektywnego uznania, co rodzi sporo kontrowersji. Nie są to jedyne istniejące definicje tego konstruktu. Interpretacja obiektu statystycznego zależy od przyjętego aparatu filozoficznego. Przedział ufności można pojmować na gruncie myśli frekwentystycznej (Neyman, Pearson, 1933), bayesowskiej albo w autorskiej koncepcji Fishera, czyli tzw. wnioskowaniu opartym na wierze (*fiducial interference*; Berger, 2003).

Drugi problem wynika bezpośrednio z konfliktu między Fisherem a Neymanem (Pearson wycofał się z dyskusji, ale zgodnie z tradycją i notacją stosowaną w tradycji, również w poniższym tekście jego nazwisko zostanie uwzględnione w nomenklaturze; za: Lenhard, 2006; Mayo, 1992). Żaden z nich nie przedstawił zbiorczej publikacji zawierającej kontrargumenty na zarzuty przeciwnika. Zamiast tego wzajemna niechęć sączyła się w kolejnych artykułach ukazujących się w różnych czasopismach naukowych przez długi czas. Szukając analogii w obecnej rzeczywistości, podążanie za tokiem rozumowania obu matematyków przypomina śledzenie wątków na forum internetowym. Osobom włączającym się do dyskusji w trakcie jej trwania trudno odnaleźć się w argumentach. Z tego względu prace porządkujące obraz sytuacji są nieocenione, np. praca Zabell (1992), która stanowi zestawienie obu podejść, czy podobnie praca Gigerenzera (1993), do których można odesłać czytelnika.

Wszystkie szczegóły konfliktu na linii Fisher–Neyman trudno zaprezentować w pojedynczym artykule, zwłaszcza że nie ograniczają się tylko do p-wartości czy podejmowania decyzji, stąd też poniższa sekcja ograniczy się do zarysu historyczno-filozoficznego procedury NHST. Więcej na ten temat można znaleźć u wyżej wspomnianych autorów: Gigerenzer (1993), Lenhard (2006) czy Zabell (1992).

Fisher jako pierwszy spośród tej trójki (Fisher, Neyman, Pearson) publikował prace z zakresu testowania hipotez wraz ze swoją autorską logiką wnioskowania indukcyjnego. W 1922 roku została wydana bardzo ważna, również dla samej statystyki matematycznej, publikacja na temat testowania hipotez – „On the mathematical foundations of theoretical statistics" . Przedstawił w niej wiele użytecznych do dzisiaj pojęć, takich jak statystyka dostateczna. Dostateczność wyjaśnia m.in., dlaczego można stosować średnią z próby zamiast analizować cały zbiór danych (Fisher, 1922). Trzy lata później, w 1925, została wydana pozycja *Statistical methods for research workers*, a w 1935 *The design of experiments*. Kolejne wydania obu książek świadczą o ich ogromnej popularności wśród naukowców.

Zdaniem Zabell (1992), Fisher zaczynał jako bayesista, który z czasem odwrócił się od konstruktu prawdopodobieństwa odwrotnego (wynikającego z twierdzenia Bayesa) i rozwinął własną koncepcję pozbawioną prawdopodobieństw apriorycznych – tzw. koncepcję prawdopodobieństwa opartego na wierze (*fiducial probability*), które pozwala na wnioskowanie od szczegółu do ogółu na zupełnie innych zasadach. Właśnie ona tłumaczy powstanie i interpretację p-wartości (Hubbard, Bayarri, 2003). Filozofię Fishera mało kto dobrze pojmuje, zwłaszcza że w trakcie życia autora ulegała istotnym zmianom (Dienes, 2008) i do dzisiaj chętnie korzysta się z jej owoców.

Najbardziej charakterystycznym elementem paradygmatu Fisherowskiego jest hipoteza zerowa, odnosząca się pierwotnie do parametru populacji (rozszerzeń na modele nieparametryczne dokonano w późniejszych latach). Na jej podstawie budowana jest statystyka testowa, czyli obiekt matematyczny wiążący próbę oraz parametr postulowany przez hipotezę zerową. Jest ona funkcją próby, czyli pewnym przekształceniem danych spełniającym wymagane kryteria, m.in. wspomnianą już dostateczność. Zakładając prawdziwość hipotezy zerowej, znany jest jej rozkład, np. w jednopróbowym teście *t*-Studenta rozkład statystyki testowej jest rozkładem *t*-Studenta z $n - 1$ stopniami swobody. Zebrana próba pozwala wyliczyć wartość statystyki testowej. Ten wynik oraz znajomość funkcji rozkładu umożliwia ocenę, czy uzyskany rezultat to wartość typowa w zbiorze wszystkich możliwych wartości statystyki testowej. Ocena ekstremalności wyników to właśnie

p-wartość (*p-value*). W poniższych akapitach znajduje się jej matematyczna definicja.

Im bliżej jedności sięga p-wartość, tym bardziej typowa jest wartość statystyki testowej. Małe wartości p-wartości oznaczają, że obliczona wartość statystyki testowej należy do rzadko spotykanych pod warunkiem prawdziwości hipotezy zerowej. Kryterium istotności wyników, nazywane poziomem istotności (alfa, α), posiada zaproponowaną przez Fishera wartość referencyjną równą 0,05. Według niego: „jest to normalne i wygodne dla badaczy, aby przyjąć 5% jako standardowy poziom istotności, w tym sensie, że są oni gotowi ignorować wszystkie rezultaty, które nie osiągnęły tego standardu i odrzucić je" (Fisher, 1955, s. 13).

Fisher wskazuje badaczowi, jak należy rozumieć te rezultaty, które są istotne statystycznie. Nie działa to w drugą stronę. Nie wiadomo, czy nieistotne statystycznie wyniki upoważniają do stanowczego stwierdzenia, że zjawisko nie zachodzi. Jak piszą Halpin i Stam (2006), interpretacja nieistotnych statystycznie wyników w testach istotności Fishera jest niejednoznaczna. Można odnieść wrażenie, że chodzi o wybiegi językowe. Fisher stał na stanowisku, że hipotezy zerowej nie można *udowodnić* (*prove*) ani *ustanowić* (*establish*), ale można ją obalić na drodze eksperymentu (Fisher, 1971): „można powiedzieć, że każdy eksperyment istnieje tylko po to, aby stworzyć szansę obalenia hipotezy zerowej" (s. 16). Fisher oburzał się, słysząc błędne stwierdzenia badaczy mówiących o hipotezie zerowej jako „zaakceptowanej-ale-fałszywej" (*accepted when false*), kiedy test pokazał wynik nieistotny statystycznie (1955). Hipoteza zerowa dla Fishera może być „co najwyżej potwierdzona (*confirmed*) albo wzmocniona (*strenghtened*)" (s. 73).

Brak jednoznacznych interpretacji statystycznej nieistotności wyników sprowadził naukowców nie statystyków na manowce. Pod tym względem atrakcyjniejsza wydaje się być myśl Neymana i Pearsona, która jasno określa reguły postępowania. W tym miejscu dobrze uświadomić sobie odrębność centralnych elementów dla obu podejść. Test istotności to obiekt charakterystyczny dla paradygmatu Fishera. Natomiast podejmowanie decyzji w związku z otrzymanym wynikiem jest zagadnieniem, jakim zajmowało się dwóch innych matematyków – Jerzy Spława-Neyman, doktor matematyki z Polski, oraz Egon S. Pearson, syn Karla Pearsona, znanego m.in. z wprowadzenia pojęcia współczynnika r Pearsona. Swoje prace nad testowaniem hipotez prowadzili jeszcze w latach dwudziestych XX wieku. Podejście prezentowane przez powyższy duet (odtąd w skrócie N–P) jest przez wielu uważane za bardziej eleganckie matematycznie z uwagi na udowodniony lemat[1], nazywany lematem Neymana–Pearsona, który podaje sposób konstruowania testu specjalnego typu (w terminologii matematycznej *jednostajnie najmocniejszego*). Publikacja tego lematu i zarazem kulminacja współpracy między matematykami przypadła na rok 1933. Potem ich drogi się rozeszły.

Punktem wyjścia w paradygmacie N–P jest istnienie dwóch równoważnie traktowanych hipotez. Współcześnie nazywane są one *zerową* i *alternatywną*. Dla Neymana były to dwie *alternatywne hipotezy*. Dzięki lematowi N–P badacz może podjąć decyzję, jak postąpić w kwestii wyboru hipotezy, jednak nie jest w tym nieomylny. Może odrzucić prawdziwą w rzeczywistości hipotezę zerową (i przyjąć fałszywą hipotezę alternatywną) – jest to błąd I rodzaju (*false positive*), zwany alfa. Może też przyjąć fałszywą w rzeczywistości hipotezę zerową (i zignorować prawdziwą hipotezę alternatywną) – jest to błąd II rodzaju (*false negative*), beta. Błąd II rodzaju to błędne stwierdzenie braku związku między zmiennymi.

Zwykle alfa wynosi 0,05. Wartość ta do złudzenia przypomina poziom istotności Fishera, wyznaczony również na 0,05. Zakładając, że hipoteza zerowa jest prawdziwa, badacz, który wykonuje nieskończenie wiele powtórzeń eksperymentu, będzie mylić się i błędnie odrzucać prawdziwą hipotezę zerową w 5% przypadków.

W celu podjęcia decyzji, którą hipotezę należy wybrać, trzeba zająć się nierównością uwzględniającą stosunek dwóch funkcji wiarogodności, tzw. iloraz wiarygodności (*likelihood ratio*). Funkcja wiarygodności (*likelihood function*) jest to funkcja wiążąca gęstość próby z parametrami zawartymi w hipotezie (zerowej albo alternatywnej). W lemacie N–P iloraz wiarygodności wykorzystuje się do podziału zbioru możliwych wyników statystyki testowej na dwa rozłączne podzbiory – zbiór krytyczny i zbiór przyjęć (te dwa pojęcia niekiedy są podawane na kursach ze statystyki). Jeśli wartość statystyki testowej, obliczona na podstawie zebranych danych, leży w zbiorze wartości krytycznych, to hipoteza zerowa zostanie odrzucona na korzyść hipotezy alternatywnej. W przeciwnym przypadku podejmujemy decyzję o przyjęciu hipotezy zerowej.

Podejście Neymana i Pearsona należy do teorii podejmowania decyzji – badacz podejmuje decyzję, co do przyjęcia lub odrzucenia hipotezy, lecz na podstawie konkretnych wyników badań nie może wiedzieć, która hipoteza jest błędna, wie natomiast, ile razy będzie popełniał pomyłkę (błąd I rodzaju), gdyby przeprowadzał badanie nieskończoną liczbę razy. Ten brak informacji o prawdziwości konkretnej hipotezy podkreślali sami autorzy paradygmatu: „Jesteśmy skłonni sądzić, że jeśli chodzi o konkretną hipotezę, żaden test oparty na teorii prawdopodobieństwa nie jest w stanie dostarczyć żadnych dowodów prawdziwości czy fałszywości tej hipotezy" (Neyman, Pearson, 1933, s. 291–292).

[1] Lemat to twierdzenie pomocnicze.

Mimo że pierwotnie prace Neymana i Pearsona miały być uzupełnieniem podejścia Fishera, to zostały one odrzucone i ostro skrytykowane:

> Istnieje wyraźna różnica w logicznym sposobie myślenia i powstała ona kiedy Neyman, myśląc, że poprawia i ulepsza moje wcześniejsze prace nad testami istotności, w zasadzie zreinterpretował je w odniesieniu do tego technologicznego i komercyjnego aparatu, który znany jest pod nazwą procedury akceptacji. (Fisher, 1955, s. 69)

Fisher nie zaakceptował alternatywnych koncepcji. Kwestia testowania hipotez nie była zresztą pierwszą potyczką – różnica zdań zaczęła się od przedziałów ufności i szybko przeszła do wzajemnej niechęci, której nie udało się uniknąć do końca życia. Nuzzo (2014) twierdzi, że świat naukowy, zmęczony ustawiczną walką między rywalami, a będący w potrzebie przeprowadzania analiz, wziął sprawy w swoje ręce, pisząc podręczniki bez rzetelnej i dogłębnej wiedzy, zarazem ryzykując poważnymi błędami. Tezę tę potwierdzają Halpin i Stam (2006), pisząc, że autorzy podręczników o statystyce przeznaczonych dla psychologów czuli się w obowiązku, aby pogodzić oba te podejścia. Jeszcze w latach czterdziestych ukazała się książka autorstwa Lindquista pt. *Statistical analysis in educational research*, której obecnie przypisuje się popularyzację NHST (Halpin, Stam, 2006).

Według filozofa Lenharda (2006) podejścia zawarte w tej procedurze nie są kompatybilne ze względu na to, że: „obie strony miały odmienne zdanie na temat funkcji modeli matematycznych i roli modelowania w statystycznym wnioskowaniu" (s. 71) – zrozumienie tego stwierdzenia wymagałoby sięgnięcia do annałów filozofii. Na potrzeby niniejszego artykułu wystarczy proste wyjaśnienie Royalla (1997): „Najważniejsza różnica między N–P a Fisherem tkwi w celu. Te pierwsze są regułami wyboru między alternatywnymi czynnościami, testy istotności mają mierzyć dowód" (s. 64).

Zgodnie z powyższym, test Neymana–Pearsona pozwala podjąć decyzję, co zrobić z otrzymanym wynikiem, natomiast celem testu istotności jest móc wyciągnąć wnioski o konkretnym wyniku badania w kontekście posiadania dowodu przeciwko hipotezie zerowej.

Różnicę w podejściach odzwierciedla również nomenklatura – podejście Neymana–Pearsona, zwane *fixed – alpha/fixed – level approach* jest podejściem frekwentystycznym. Według Neymana było ono „indukcyjnym postępowaniem" (*inductive behaviour*) w opozycji do „indukcyjnego wnioskowania" (*inductive inference/reasoning*) Fishera, którego podejście, nazywane *p-value approach*, nie jest podejściem ani frekwentystycznym, ani bayesowskim – jak pisze Dienes (2008).

Ciekawe rozróżnienie proponuje Lew (2012), pisząc, że podejście Fishera jest „lokalne", ponieważ wynik testu istotności odnosi się do konkretnej hipotezy i konkretnego badania. Podejście Neymana–Pearsona określa „globalnym" ze względu na to, że w tym paradygmacie nie można stwierdzić prawdziwości hipotezy w pojedynczym badaniu. Ważniejszy od tego jest długoterminowy błąd I rodzaju oraz reguły decyzyjne z nim związane (Lew, 2012). W celu uzupełnienia obrazu sytuacji poniżej została przedstawiona procedura NHST w formie wykorzystywanej do dzisiaj:

1. Ustal poziom istotności alfa = 0,05.
2. Sformułuj dwie hipotezy testowe: hipotezę zerową (H_0) oraz hipotezę alternatywną (H_1).
3. Na podstawie danych oblicz statystykę testową.
4. Znając rozkład statystyki testowej, oblicz p-wartość oraz:

- jeśli p-wartość jest mniejsza niż alfa, wtedy można stwierdzić:

„Podejmuję decyzję o odrzuceniu hipotezy zerowej na korzyść alternatywnej";

- jeśli p-wartość jest większa niż alfa, to można stwierdzić:

„Nie ma podstaw do odrzucenia hipotezy zerowej" (względnie „na korzyść hipotezy alternatywnej"). Gigerenzer, niemiecki psycholog, ochrzcił procedurę NHST mianem „hybrydy" (1993). *Inny słownik języka polskiego* definiuje hybrydę jako „coś, co składa się z różnych elementów, często do siebie niepasujących" (Bańko, 2000, s. 519). Zgodnie z tą definicją, analizując powyższe kroki procedury NHST, można wyszczególnić, które z nich należą do Fishera, a które do Neymana–Pearsona.

Z koncepcji Neymana–Pearsona zaczerpnięty jest koncept decyzji, hipoteza alternatywna, błędy I i II rodzaju. Natomiast p-wartość jest elementem jedynie podejścia Fishera. W wielu badaniach nie oblicza się prawdopodobieństwa popełnienia błędu II rodzaju czy szerzej – nie oblicza się mocy. Są to konsekwencje przyjęcia Fisherowskiego sposobu myślenia, gdzie hipoteza alternatywna nie jest potrzebna.

Dla porównania w załączniku B czytelnik może porównać sposób testowania hipotezy według Fishera oraz według Neymana–Pearsona.

Poniższe dwie sekcje będą poświęcone p-wartości, która występuje w NHST, ale przede wszystkim jest elementem podejścia Fishera. Większość zarzutów skierowanych przeciwko NHST jest zarzutami skierowanymi właśnie przeciwko p-wartości.

Czym *p*-wartość nie jest? Błędne przekonania

W środowisku naukowym ciągle krążą błędne przekonania o p-wartości – czym naprawdę jest, prawdopodobieństwo czego (jakiego zdarzenia) kryje się za tym terminem? Zanim czytelnik przejdzie dalej, niech spróbuje samodzielnie wykonać poniższe zadanie:

> Wynik wynosi $t(38) = 2,7$; $p = 0,01$. Proszę zdecydować, które z poniższych twierdzeń są prawdziwe, a które są fałszywe:
> (i) całkowicie sfalsyfikowano hipotezę zerową (nie ma różnic między średnimi w populacjach);
> (ii) znaleziono prawdopodobieństwo prawdziwości hipotezy zerowej;
> (iii) udowodniono hipotezę naukową (o istnieniu różnicy między średnimi w populacjach);
> (iv) można wydedukować prawdopodobieństwo prawdziwości hipotezy zerowej;
> (v) znane jest prawdopodobieństwo popełnienia błędu, jeśli zdecydowano, by odrzucić hipotezę zerową;
> (vi) uzyskano wiarygodne odkrycie eksperymentalne w tym sensie, że hipotetycznie, jeśli eksperyment byłby przeprowadzany olbrzymią liczbę razy, to otrzymano by istotny wynik w 99%.
>
> (Oakes, 1986, za: Dienes, 2008)

Powyższe zadanie przedstawiono 70 osobom (naukowcom, nauczycielom akademickim z co najmniej dwuletnim stażem), które można było podejrzewać o poziom wiedzy umożliwiający poprawną odpowiedź, czyli zakwalifikowanie każdej z opcji jako fałszywą. Tymczasem tylko dwie osoby udzieliły prawidłowej odpowiedzi (Dienes, 2008). 60% odpowiedzi wskazało na niepoprawną odpowiedź vi (Gigerenzer, 1993). Niestety, żadna z powyższych nie jest prawdziwa.

Przekonanie, że p-wartość to prawdopodobieństwo prawdziwości hipotezy zerowej jest powszechne. Jednakże p-wartość jest prawdopodobieństwem specjalnego rodzaju – prawdopodobieństwem warunkowym[2], definiowanym jako prawdopodobieństwo otrzymania wyników (czy wartości statystyki testowej) co najmniej tak ekstremalnych, jak uzyskane w badaniu, ale pod warunkiem założenia prawdziwości hipotezy zerowej. Wynika stąd, że w celu oceny typowości uzyskanego wyniku zakłada się prawdziwość hipotezy zerowej.

Zapis matematyczny p-wartości przyjmuje postać: $P(T(X) > x|H_0)$, gdzie T jest statystyką testową. Odnosi się ona do ogonów gęstości (krańcowych części tej krzywej) rozkładu statystyki testowej. Dla ułatwienia zapamiętania można go skrócić do bardziej intuicyjnej formy: $P(\text{dane}|H_0)$.

[2] P(A|B) czytane jest jako prawdopodobieństwo zdarzenia A pod warunkiem zdarzenia B, zobacz: załącznik A.

Jak słusznie pisze Gigerenzer (1993), ani teoria Fishera, ani teoria Neymana–Pearsona, ani hybrydowe połącznie obu nie podaje tego, czego życzyłby sobie badacz – prawdopodobieństwa prawdziwości hipotezy zerowej pod warunkiem uzyskania danych ($p(H_0|\text{dane})$), zamiast tego p-wartość podaje $p(\text{dane}|H_0)$.

Prawdopodobieństwa warunkowe nie są odwrotne względem siebie. Nie można stwierdzić konkretnej wartości P(B|A), posiadając jedynie wartość P(A|B), lub też stwierdzić, że są one zbieżne (zob. załącznik A). Prawdopodobieństwo P(A|B) nie jest równe prawdopodobieństwu P(B|A), ani też nie musi być proporcjonalne. Prawdopodobieństwa te, P(A|B) i P(B|A), są związane równaniem Bayesa:

$$P(B|A) = \frac{P(A|B) * P(B)}{P(A)}.$$

Prawdopodobieństwo zdarzenia B pod warunkiem zajścia zdarzenia A może zostać obliczone, jeśli znane są prawdopodobieństwo zdarzenia A pod warunkiem zajścia zdarzenia B oraz prawdopodobieństwo zdarzenia A, P(A), i prawdopodobieństwo zdarzenia B, P(B). Znaczenie ma kolejność warunkowania (czy na zdarzenie A, czy na zdarzenie B) – bez znajomości prawdopodobieństwa zdarzenia, na które warunkuje się (np. P(B)), nie można wydedukować prawdopodobieństwa zdarzenia warunkowanego A. Nie można stąd obliczyć P(A|B) na samej podstawie P(B|A), jak również twierdzić, że są zbliżone.

Poniższy przykład dobrze ilustruje tę sytuację. Prawdopodobieństwo, że trafiła się karta koloru trefl, jeśli wiadomo, że wylosowano króla P(trefl|król) wynosi 0,25. Z drugiej strony prawdopodobieństwo wylosowania króla, jeśli wiadomo, że karta jest treflowa P(król|trefl) jest równe 0,08.

W związku z nierównoważnością prawdopodobieństw $P(D|H_0)$, czyli p-wartości, oraz $P(H_0|D)$ nie można wnioskować o prawdziwości czy też fałszywości hipotezy zerowej, jeśli wcześniej założyło się prawdziwość H_0. Również nie można twierdzić, że p-wartość jest dobrym przybliżeniem prawdopodobieństwa prawdziwości hipotezy zerowej pod warunkiem danych.

Nieprawdziwy jest taki wniosek: „jeśli p-wartość = 0,03, to oznacza to, że hipoteza zerowa ma tylko 3% szans na bycie prawdziwą". Skoro już wcześniej założono prawdziwość hipotezy zerowej, co zawiera się w definicji p-wartości, to poprzednie zdanie nie ma sensu – nie obliczono prawdopodobieństwa hipotezy zerowej (które zresztą w nurcie frekwentystycznym nie jest możliwe). P-wartość podaje szansę uzyskania podobnego lub bardziej ekstremalnego wyniku przy konkretnym założeniu logicznego statusu hipotezy zerowej (*prawda*).

Statystyczna istotność bywa również rozpatrywana w kategoriach istotności rzeczywistej, np. $t(40) = 1{,}2$; $p = 0{,}103$ oznacza *wynik nieistotny statystycznie*, ale nie oznacza braku różnic pomiędzy grupami. Statystyczna (nie) istotność nie jest tożsama z rzeczywistą (nie)istotnością. Oznacza jedynie, że wynik osiągnął (bądź nie) pewne arbitralne kryterium 0,05. Przy ocenie istotności rzeczywistej można posłużyć się pewną miarą, tzw. wielkością efektu (*effect size*) – choć nierzadkie są badania, w których jest ona bardzo mała przy jednoczesnej statystycznej istotności.

Statystyczna nieistotność nie może być traktowana jako potwierdzenie hipotezy zerowej. Ma to związek z asymetrycznym zachowaniem się p-wartości względem hipotezy zerowej. O ile wynik istotny statystycznie w paradygmacie Fishera świadczy przeciwko hipotezie zerowej, o tyle wynik nieistotny statystycznie nie niesie żadnej informacji popierającej hipotezę zerową. Zdarzają się fatalne w skutkach decyzje podjęte na podstawie nieistotności wyników. Na wspomnienie zasługuje przypadek paroksytyny – leku przeciwdepresyjnego, który dopuszczono do użytku, ponieważ wyniki badań nad podwyższonym ryzykiem samobójstwa u dzieci pod wpływem tego środka były nieistotne statystycznie (Dienes, 2016).

Wobec powyższego, niuansem językowym wydaje się być następujące stwierdzenie: „hipoteza zerowa mówi, że nie ma istotnych statystycznie różnic w dwóch populacjach", obecne w wypowiedziach psychologów i wynikające być może z niezrozumienia filozoficznego aspektu testowania hipotez. Dienes (2008) skrytykował takie sformułowanie, argumentując, że „istotność nie jest własnością populacji" (s. 72) – hipoteza zerowa wyraża, że nie ma różnic w dwóch populacjach.

W artykułach wciąż używa się zapisu istotnych statystycznie wyników za pomocą *gwiazdek*, czyli tzw. asterysków *. Im więcej gwiazdek, tym bardziej istotny wynik $p < 0{,}01$*; $p < 0{,}05$**; $p < 0{,}001$***. Niestety zapis ten jest nieprawidłowy i coraz więcej czasopism wymaga, aby podawać konkretne wartości p-wartości. Jest to jedno z zaleceń szóstej edycji Amerykańskiego Towarzystwa Psychologicznego (APA, 2011). Podawanie p-wartości w formie nierówności jest znów konsekwencją połączenia paradygmatów Fishera i Neymana–Pearsona. Fisher opowiadał się za tym, aby naukowiec podawał dokładną wartość p-wartości (Gigerenzer, 2004). W podejściu N–P wartość statystyki testowej albo wpada w zbiór wartości krytycznych (powodujących odrzucenie hipotezy zerowej), albo w zbiór przyjęć. Badacza nie interesuje, jak bardzo jest ta wartość oddalona od granicy między tymi dwoma zbiorami. Z drugiej strony, zgodnie z koncepcją Fishera, p-wartość jest siłą dowodu przeciwko hipotezie zerowej (Fisher, 1958, za: Royall, 1997), zatem oprócz faktu znajdowania się po lewej (< 0,05) lub prawej (> 0,05) stronie poziomu istotności 0,05, ważna jest konkretna wartość liczbowa.

Dwa unikalne dla obu koncepcji parametry (p-wartość i alfa α) również są traktowane zamiennie. Jest to błąd (Hubbard, Bayarri, 2003). P-wartość nie ma nic wspólnego z alfa rozumianym jako prawdopodobieństwo popełnienia błędu pierwszego rodzaju. Błędne są stwierdzenia typu: „*p* jest mniejsze niż prawdopodobieństwo popełnienia błędu I rodzaju". Jest to pomieszanie p-wartości, pochodzącej z podejścia Fishera, i alfa, odnoszącej się do błędu I rodzaju z podejścia Neymana–Pearsona.

Poziom istotności alfa w koncepcji Fishera jest pozbawiony częstościowej interpretacji, jaką posiada alfa w paradygmacie Neymana–Pearsona (równe prawdopodobieństwu popełnienia błędu I rodzaju w długoterminowym wykonywaniu eksperymentu). Fisher nie zakłada nieskończonej liczby powtórzeń badania, stąd nie można mówić o częstości popełniania pomyłki w kontekście p-wartości. W jego filozofii poziomu istotności nie można interpretować w terminach częstości, co on sam wyraża słowami:

> W ostatnim czasie, jedno częste wyjaśnienie testów istotności, które jest odpowiedzialne za sprowadzanie na manowce czytelników matematycznych poprzez twierdzenie czegoś, co nie jest ogólnie prawdą, że poziom istotności musi być równy częstotliwości z jaką hipoteza jest odrzucana w powtarzalnym próbowaniu z ustalonej populacji dopuszczonej przez hipotezę. Ten natrętny aksjomat, obcy rozumowaniu testów istotności, jest w zasadzie prawdziwą przeszkodą w postępie. (Fisher, 1945, s. 130, za: Hubbard, Bayarri, 2003 lub Pearson, 1966, s. 172)

Inną konsekwencją pomieszania błędu I rodzaju oraz poziomu istotności jest tzw. postulat alfa (Royall, 1997), czyli podział zakresu możliwych p-wartości na przedziały o mniejszej lub większej sile dowodu przeciwko hipotezie zerowej. Na przykład p-wartość między 0,01 a 0,05 oznacza *wynik istotny statystycznie*, podczas gdy p-wartość sięgająca poniżej 0,01 oznacza – zdaniem badacza – *wynik bardzo istotny statystycznie*. Wynik bliski statystycznej istotności to taki, którego p-wartość mieści się w przedziale 0,1 do 0,05. Wynik jest albo statystycznie istotny, albo nie. Być może badacze dzielą p-wartości na obszary o różnej istotność, aby ułatwić sobie porównywanie wyników z kilku badań. Nie są jednak świadomi faktu, że p-wartości nie można między sobą porównywać (Gelman, 2012).

Zarzuty stawiane p-wartości

Błędy logiczne w NHST

Wielu zarzuca, że NHST (dotyczy to części opartej na testach istotności) jest błędna logicznie z uwagi na interpretację p-wartości (Cohen, 1994; Dienes, 2008). Fisher pisał, że p-wartość jest *siłą* dowodu przeciwko hipotezie

zerowej (*strength of evidence*). Wynik istotny statystycznie oznacza, że: „albo zdarzył się wynik ekstremalny albo teoria nie jest prawdziwa" (Fisher, 1959, za: Royall, 1997). Słowo *teoria* jest tutaj użyte w znaczeniu hipoteza zerowa. Ta alternatywa rozłączna jest znana pod nazwą dysjunkcji Fishera. Można ją przeformułować do postaci sylogizmu:

(Sylogizm A)
Jeśli H_0 jest prawdziwa, to prawdopodobnie wynik ekstremalny nie pojawiłby się.
Wystąpił wynik ekstremalny.
Stąd H_0 jest prawdopodobnie nieprawdziwa.

Sylogizm A jest identyczny (z wyjątkiem jednego wyrazu) z poniższym sylogizmem B:

(Sylogizm B)
Jeśli H_0 jest prawdziwa, to wynik ekstremalny nie pojawiłby się.(*)
Wystąpił wynik ekstremalny.
Stąd H_0 jest nieprawdziwa.

Postać pierwszego z trzech zdań (oznaczonego gwiazdką *, bezpośrednio powyżej) w sylogizmie B jest implikacją: „jeśli p, to q". Zdanie p to zdanie: „H_0 jest prawdziwa". Zdanie q to zdanie „wynik ekstremalny nie pojawiłby się". Prawo kontrapozycji, „jeśli nie q, to nie p", upoważnia do wniosku, że hipoteza zerowa jest nieprawdziwa (nie p), skoro wynik ekstremalny pojawił się (nie q).

Różnica między sylogizmami A i B polega na obecności elementu probabilistycznego. Pierwszy sylogizm zawiera słowo *prawdopodobnie*. Logika arystotelesowska nie działa w takich przypadkach.W celu obrony dysjunkcji Fishera jeden z badaczy przeformułował ją następująco (Lew, 2013):

Ekstremalne p-wartości z próby losowej są rzadkie pod warunkiem hipotezy zerowej.
Ekstremalna p-wartość została zaobserwowana.
Stąd hipoteza zerowa jest prawdopodobnie fałszywa.

Przy takim rozumowaniu ostatnie zdanie nie rozwiązuje problemu – nadal nie wiadomo jaki status logiczny ma hipoteza zerowa.

Wnioskowanie dedukcyjne jest możliwe, jeśli ani wniosek, ani przesłanka nie są probabilistyczne (Dienes, 2008).

Fisher miał prawo sądzić, że p-wartość jest siłą dowodu przeciwko hipotezie zerowej, ale jest to zupełnie inna rzecz niż prawdopodobieństwo otrzymania takiego dowodu (Dienes, 2008) – p-wartość nie jest miarą siły dowodu, w związku z czym nie dostarcza dowodów przeciwko hipotezie zerowej (Wagenmakers, 2007). Gdyby tak było, wówczas p-wartość pochodząca z dwóch prób o dwóch różnych liczebnościach miałaby taką samą wagę – jest to założenie znane pod nazwą *postulat p* (*p-postulate*; Wagenmakers, 2007).

Przykładowo: w eksperymencie K dla próby o liczności $N = 30$ p-wartość wynosi 0,03. Z kolei w eksperymencie L dla próby o liczności $N = 60$ p-wartość również wynosi 0,03. Czy p-wartość w obu badaniach ma taką samą moc? Są zwolennicy obu opcji. Można być przekonanym, że im większa próba, tym mocniejszy jest dowód. Tak samo można uważać, że odrzucenie hipotezy zerowej na mniej licznej próbie dostarcza dowodu bardziej wiarygodnego. Analiza bayesowska sugeruje, że rację mają ci drudzy – jeśli w wyniku przeprowadzenia dwóch eksperymentów uzyskujemy tą samą p-wartość w obu przy różnych liczebnościach obu próbek, eksperyment o mniej licznej próbie dostarcza mocniejszego dowodu przeciwko hipotezie zerowej (Wagenmakers, 2007; Royall, 1997).

Zależność od wielkości próby

Wielu krytyków NHST słusznie podkreśla zależność p-wartości od wielkości próby. Na przykładzie badań współprowadzonych przez autorkę tekstu nad adaptacją *Kwestionariusza kodów moralnych* (Jarmakowski-Kostrzanowski, Jarmakowska-Kostrzanowska, 2014) zostanie pokazane, jak zmienia się p-wartość w zależności od liczby elementów wchodzących do obliczenia wartości współczynnika korelacji między kodami Opieka i Sprawiedliwość, rozpoczynając od próby 10-elementowej i stopniowo dobierając kolejne elementy ze zbioru.

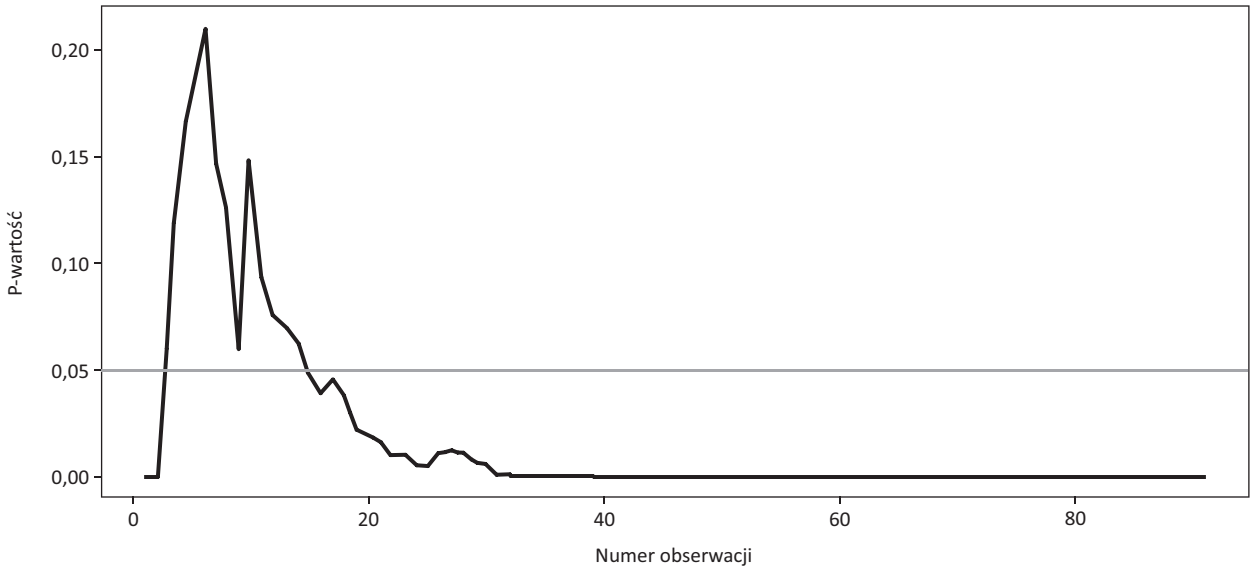
Rysunek 1 pokazuje, że wartość p-wartości szybko spada poniżej wartości progowej 0,05. W celu rozwiania wszelkich wątpliwości przeprowadzona została dodatkowa analiza. Tym razem zwiększano zarówno rozmiar próbki, jak i elementy wchodzące do niej losowano z całego zbioru. Rozpoczynając od 10-elementowej próby, do której wylosowano 10 elementów (np. mogły wejść obserwacje z początku zbioru, środka i końca), obliczano współczynnik korelacji, pobierano p-wartość, po czym zwracano te 10 elementów z powrotem do próby. Wówczas algorytm losował 11 elementów z próby. Procedura została prowadzona aż do użycia pełnego zbioru danych. Wyniki ilustruje rysunek 2.

Początkowe fluktuacje w p-wartości jeszcze przed 40-elementową próbą zostają wygaszone poniżej 0,05.

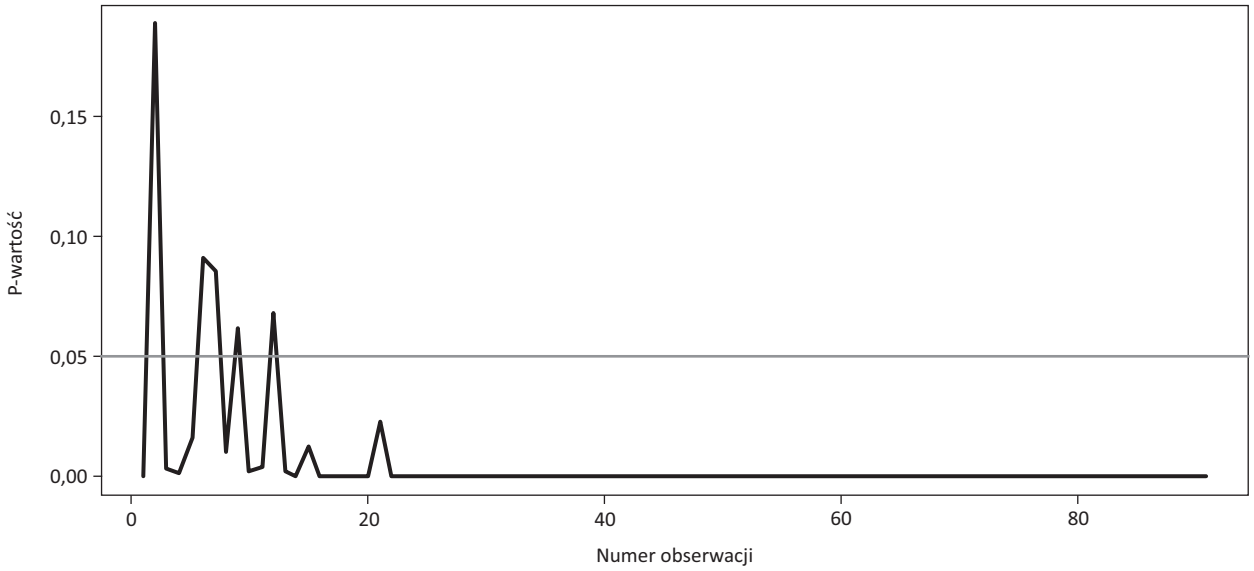
Te dwa przykłady pokazują, jak wielkość próby wpływa na istotność statystyczną. Równocześnie niestety dzięki tym manipulacjom można osiągnąć sukces publikacyjny.

Zasady stopu (*stopping rules*)

Problemem pokrewnym do problemu zależności p-wartości od wielkości próby jest zależność p-wartości od zasady stopu (*stopping rules*). Zasada stopu to warunki, pod jakimi badacz zaprzestaje zbierania danych (Dienes, 2008). Bardzo często w psychologii badania prowadzone są



Rysunek 1. Wartość p-wartości w zależności od liczby obserwacji.
Źródło: rysunki 1–4 – opracowanie własne.



Rysunek 2. Wartość p-wartości w zależności od liczby obserwacji losowanych z próby.

dotąd, dopóki są ochotnicy do wzięcia udziału – rozdawane są kwestionariusze osobom znajomym, bezpośrednio lub drogą mailową, zapraszani są studenci, współpracownicy. Ten sposób gromadzenia danych tworzy przypadkową próbę złożoną z dostępnych dla badacza osób (tzw. *convenience sample*). Istnieje przy tym pokusa, aby zbierać dane do momentu uzyskania statystycznie istotnego wyniku. Takie postępowanie wywiera bardzo duży wpływ na wartość p-wartości.

Przypuśćmy, że psychologowie A i B prowadzą badania, których celem jest sprawdzenie powszechności zjawiska X (dowolne zjawisko). Jeśli jest ono przypadkowe, to

prawdopodobieństwo jego występowania θ, wynosi $\frac{1}{2}$. Hipoteza zerowa w tym przypadku wynosi H_0: $\theta = \frac{1}{2}$, przeciwko hipotezie alternatywnej H_1: $\theta \neq \frac{1}{2}$.

Psycholog A z góry ustalił liczebność próby. Z pewnych względów uważa, że powinien przebadać 15 osób. Po wykonaniu badania odkrywa, że 11 badanych ma własność X. Dalej oblicza p-wartość, czyli prawdopodobieństwo otrzymania takiego lub bardziej ekstremalnego wyniku, przy założeniu przypadkowego występowania zjawiska X i otrzymuje wynik:

$$\text{p-wartość}_{\text{psycholog A}} = P(\text{dane}|\theta = \frac{1}{2}) = 0{,}118$$

Psycholog B ma inny plan – będzie sprawdzał dotąd, dopóki nie spotka jedenastej osoby, która doświadcza zjawiska X. Mogłaby być to dowolna osoba, od jedenastej do, być może, nieskończoności. Tymczasem tak się złożyło, że jedenasta osoba z X jest piętnastą z kolei przebadaną osobą. Czy p-wartość otrzymana w wyniku takiego zbierania danych (próbkowania) jest identyczna z p-wartością psychologa A? Odpowiedź brzmi: nie.

Rozkład statystyki testowej w badaniu psychologa A jest rozkładem dwumianowym – zaplanowano dokładnie liczebność próby: 15 osób. Dla każdej z nich szansa na zjawisko X wynosi $\frac{1}{2}$. Prawdopodobieństwo, że żadna z nich nie ma zjawiska X wynosi $(\frac{1}{2})^{15}$.

Z kolei pomysł psychologa B na sprawdzanie badanych dopóki nie znajdzie jedenastej osoby wymusza zmianę planu próbkowania, co pociąga za sobą zmianę rozkładu statystyki testowej na ujemny dwumianowy. Psycholog B w obliczeniu p-wartości musi uwzględnić niepewność co do numeru osoby doświadczającej X. Stąd p-wartość psychologa B wynosi:

$$\text{p-wartość}_{\text{psycholog B}} = P(\text{dane}|\theta = \frac{1}{2}) = 0{,}034$$

Pozornie nieznacząca różnica w zasadach zaprzestania zbierania danych powoduje, że psycholog A stwierdzi losowość zjawiska X w populacji, natomiast psycholog B powie, że istnieje pewien wzorzec występowania tego samego zjawiska.

Podejście Neymana–Pearsona jasno wyznacza regułę stopu – projektując badanie należy określić pewien parametr, zwany mocą testu. Moc testu oznacza zdolność danego testu do niepopełnienia błędu II rodzaju, a ujmując to nieco prościej – do rozpoznania fałszywej hipotezy zerowej i odrzucenia jej na korzyść hipotezy alternatywnej.

$$\text{moc} = P(\text{odrzucenie } H_0|H_1 \text{ jest prawdziwa})$$

Matematycznie, moc i beta są związane prostym równaniem[3]:

$$\text{moc} = 1 - \text{beta}$$

Moc testu równa 0,50 oznacza 50% szansę na odrzucenie fałszywej hipotezy zerowej. W takim razie ile powinna wynosić moc testu? Cohen (1992) sugerował, że rozsądny poziom mocy w badaniach powinien wynosić 0,80. Moc testu, wielkość efektu, liczność próby oraz alfa to cztery parametry, które są ze sobą związane – mając trzy z nich, można obliczyć czwartą. Przed rozpoczęciem badań zaleca się obliczyć liczbę osób badanych tak, aby badania nie były ani za słabe, ani za mocne (Ellis, 2010)[4]. Testy mogą mieć zbyt małą moc (*underpowered*) i wówczas zbyt często popełniają błąd II rodzaju, jak również mogą mieć zbyt dużą moc (*overpowered*), tzn. w bardzo dużych próbach łatwo osiągnąć statystyczną istotność (małą wartość p-wartości) przy niezbyt istotnym (w sensie praktycznym) wyniku. Ellis (2010) podaje za Fieldem i Wrightem (2006) przykład, w którym na bardzo obszernej próbie test *t*-Studenta pokazał wynik istotny statystycznie (p-wartość = 0,022), przy czym różnica średnich jest równa 0 z dokładnością do dwóch miejsc po przecinku (zob. tabela 1).

P-wartość jako statystyka

Ostatni punkt trudno potraktować jako zarzut, ponieważ jest to cecha p-wartości, o której często się nie pamięta. Z jednej strony p-wartość podaje prawdopodobieństwo otrzymania wyniku równego lub bardziej ekstremalnego od uzyskanego w badaniu. Technicznie rzecz ujmując, polega to na policzeniu całki, czyli obliczeniu pola pod krzywą gęstości rozkładu statystyki testowej (przypadek jednostronnego testu; zob. rysunek 3).

Tabela 1
Test istotności przy identyczności dwóch grup

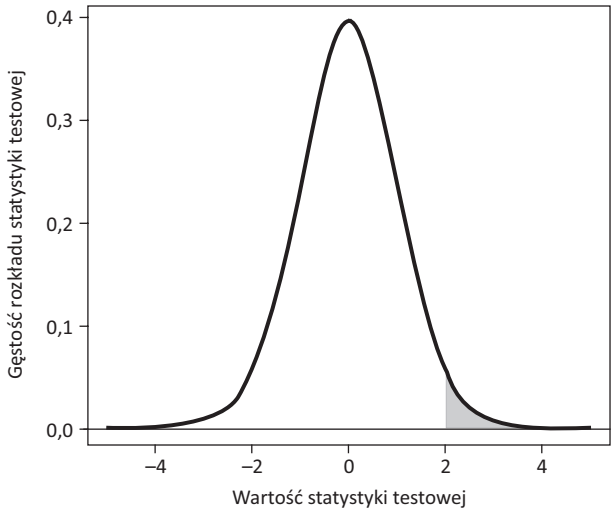
t	*df*	Istotność (p-wartość)	Różnica średnich
–2,296	999998	0,022	0,00

t – statystyka testowa; *df* – stopnie swobody.

Źródło: Field, Wright, 2006, za: Ellis, 2010.

[3] Nie jest to ani *odwrotność*, ani *przeciwieństwo*.

[4] W pakiecie R jest zaimplementowany pakiet pwr służący do obliczania mocy wielu popularnie używanych testów.



Rysunek 3. P-wartość jako obszar pod krzywą.

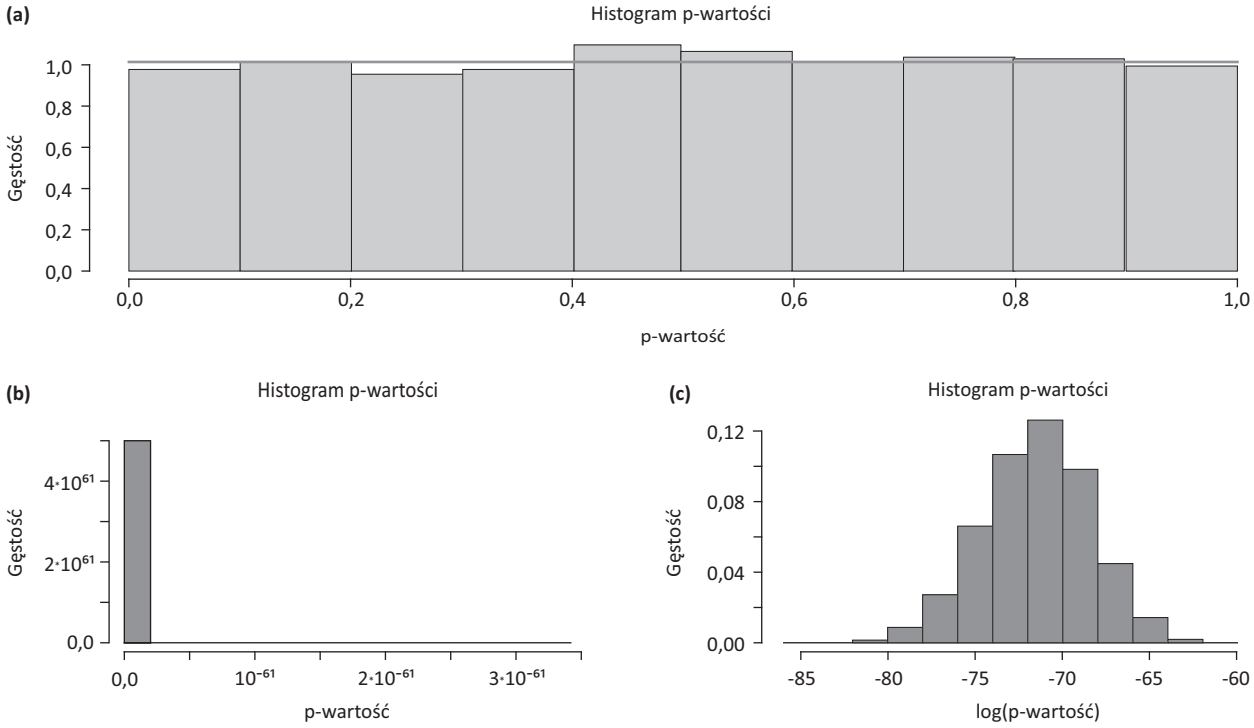
Natomiast, warto zdawać sobie sprawę z tego, że p-wartość jest statystyką – zmienną losową. W tym rozumieniu posiada ona rozkład. Jeśli zakładana przez badacza hipoteza zerowa jest w rzeczywistości prawdziwa, to rozkład p-wartości jest znanym rozkładem: standardowym jednostajnym (jednostajnym na odcinku [0,1]). Jeśli natomiast badacz myli się i hipoteza zerowa nie jest prawdziwa, to rozkład p-wartości może być dowolny (z przedziału [0,1]). Poniższy przykład ilustruje sytuację.

Przeprowadzono symulację 5000 powtórzeń jednopróbowego testu *t*-Studenta. Za każdym razem generowano 100-elementową próbę z rozkładu:

(a) normalnego o parametrach ($\mu = 5$, $\sigma = 1$),

(b) normalnego o parametrach ($\mu = 10$, $\sigma = 1$).

Badano hipotezę zerową H_0: $\mu = 5$. W pierwszym przypadku hipoteza zerowa jest prawdziwa w rzeczywistości [zmienna X jest generowana z rozkładu N(5,1)]. Rysunek 4 przedstawia histogram p-wartości pochodzący z 5000 testów. Przypomina on rozkład jednostajny.



Adnotacja. Obliczeń dokonano za pomocą programu R (2015).
Górny rysunek (a) przedstawia rozkład w postaci histogramu p-wartości w sytuacji prawdziwości hipotezy zerowej. Jak widać, rozkład ten jest bardzo zbliżony do rozkładu jednostajnego (jest bardzo spłaszczony). Dolne rysunki odpowiadają dowolnej sytuacji, w której hipoteza zerowa jest nieprawdziwa. Dolny lewy rysunek (b) przedstawia histogram takich p-wartości. Dolny prawy rysunek (c) ilustruje histogram logarytmu naturalnego z p-wartości. Potęgi p-wartości układają się w kształt dzwonu, co oznacza ich duże zróżnicowanie, uniemożliwiające stwierdzenie rozkładu jednostajnego.

Rysunek 4. Rozkłady p-wartości w zależności od prawdziwości hipotezy zerowej.

Natomiast jeśli hipoteza zerowa jest nieprawdziwa, a próby były generowane z rozkładu normalnego o innym parametrze położenia ($\mu = 10$), to histogram p-wartości pochodzących z testów ma zupełnie inną postać.

Powyższy aspekt p-wartości ma związek z zarzutami o niereprodukowalności (niereplikowalności) wyników badań. Być może czytelnikowi znany jest przypadek badań Noska, Spiesa i Motyla (2012) o czarno-białym postrzeganiu świata przez ekstremistów. Pierwsze badanie pokazało, używając nowomowy, *bardzo istotny statystycznie* rezultat, $p = 0{,}01$. Niemniej jednak zdecydowano się na powtórzenie badań. Replikacja wyników ujawniła p-wartość równą 0,59 (powtórne badania nie powtórzyły wyniku pierwowzoru). Ten przykład zaprzecza osądowi tej części badaczy, która uważa, że p-wartość oznacza prawdopodobieństwo zreplikowania otrzymanego wyniku (Gigerenzer, 1993).

Podsumowanie

Jak twierdzi wielu badaczy (m.in.: Trafimov, 2015; Cumming, 2014), procedura NHST jako skrzyżowanie dwóch podejść jest procedurą ułomną i powinna zostać zarzucona. Pod tym względem czasopismo BASP jedynie wypełniło zalecenia rekomendowane od dawna. „Testowanie istotności hipotezy zerowej jest na pewno najbardziej głupio sprowadzającą na manowce procedurą zinstytucjonalizowaną w rutynowym szkoleniu studentów" (Rozeboom, 1997, za: Trafimov, 2003, s. 526). Za to zupełnie inną sprawą jest porzucenie elementów składających się na NHST, czyli testów istotności Fishera oraz testowania hipotez w ujęciu Neymana i Pearsona. O ile ta druga metoda jest dość rzadko stosowana, o tyle testy istotności posiadają zarówno swoich zagorzałych zwolenników, jak i nieprzejednanych przeciwników. NHST powinna zostać odrzucona, ale czy testy istotności również? Zdania są podzielone.

Większość zgodzi się ze stwierdzeniem, że testy istotności ustawiają optykę badacza w czarno-białych kategoriach, dzięki kategoryzacji istotny–nieistotny statystycznie. Zauważył to m.in. Cumming (2014). W tej kwestii pytaniem kierowanym do badacza, czy nawet do osób związanych z filozofią nauki, jest to, do czego ma służyć nauka. Jedni dostrzegają naukową wartość w stwierdzeniu, że $a = b$ (średni poziom wybranej cechy jest równy w obu grupach) lub $a \neq b$ (średni poziom w grupach różni się). Właśnie takie jest pytanie, na jakie odpowiadają testy istotności (oraz NHST). Sześciu przykładów poparcia dla NHST z różnych dziedzin (w tym psychologii i fizyki) dostarcza Wainer (1999), który należy do obozu zwolenników p-wartości. Jeden z nich skutkuje wnioskiem, że dla psychologów równie ważne, jak określenie, jak bardzo, jest stwierdzenie, czy w ogóle inteligencja zmienia się w czasie.

Jednakże *statystyczna istotność* nie jest *istotnością rzeczywistą*. Nawet sam Fisher we wczesnych pracach uważał, że istotny statystycznie wynik jest tylko wynikiem godnym ponownego rozpatrzenia i sugerującym dalsze drążenie w tym temacie (Nuzzo, 2014).

Z drugiej strony psychologia, która ma ambicje stać się nauką ilościową, ma za zadanie odpowiedzieć na pytanie o wielkość badanego zjawiska. Sugerowali to Cohen (1994) i Cumming (2014). Rozwój wiedzy cierpi poprzez dychotomizację istotny–nieistotny statystycznie.

Na pewno NHST oraz testy istotności stały się bardzo wygodnym narzędziem w rękach badaczy. Najpoważniejszym problemem jest pogoń za statystyczną istotnością, która jest sprowokowana warunkami wydawniczymi (większa szansa opublikowania badań) i prowadzi do tzw. *file drawer problem*, czyli odkładania badań „do szuflady", jeśli nie udało się uzyskać rezultatów z gwiazdkami. Ponadto na poziomie samego badania pogoń za istotnością statystyczną prowadzi do odrzucania informacji, która mogłaby mieć rzeczywiste znaczenie.

Przykład: psycholog C prowadzi badanie, w którym bada, czy średni poziom cechy jest równy wartości wynikającej z teorii lub wcześniejszych badań. W tym celu gromadzi dane, rozdaje kwestionariusze lub prowadzi eksperymenty, spodziewając się, że rozkład zmiennej zależnej jest normalny. Po zebraniu wyników i przeprowadzeniu wstępnych analiz histogram pokazuje, że rozkład nie jest normalny – jest wybrzuszony w dwóch miejscach. Stąd psycholog C odwołuje się do *centralnych twierdzeń granicznych* (CTG) albo usuwa odstające obserwacje, albo przeprowadza transformacje danych tak, że ostatecznie test normalności wypada pozytywnie. Psychologowi C udaje się uzyskać wynik, że średni poziom cechy jest wyższy. Po przeprowadzeniu dalszych analiz wyniki umieszczone zostają w sekcji *Analiza* w artykule. Artykuł prawdopodobnie przejdzie pomyślnie recenzję. Poniekąd została zaprzepaszczona potencjalnie istotna informacja o nietypowym rozkładzie danych. Na tym przykładzie widać, że pogoń za statystyczną istotnością powoduje utratę cennych informacji i powstrzymuje badacza przed przyjrzeniem się swoim danym.

O ile sama procedura NHST powinna zostać zarzucona, o tyle wydaje się, że kwestia odrzucenia samych testów istotności sprowadza się do problemu, czy można wykorzystywać takie narzędzie, które daje odpowiedź na pytanie zupełnie inne niż postawione. Dla przypomnienia – p-wartość to prawdopodobieństwo otrzymania takiego wyniku statystyki testowej, jak uzyskany (lub bardziej ekstremalnego) pod warunkiem założenia prawdziwości hipotezy zerowej. Badacza interesuje prawdopodobieństwo prawdziwości hipotezy zerowej przy takim wyniku jak uzyskany.

Błędem byłoby traktowanie odpowiedzi na inne pytanie jako tej właściwej. Dopóki zdajemy sobie sprawę z tego, że dostajemy odpowiedź na pytanie, którego nie zadaliśmy, testy istotności mogłyby być drugorzędnym, jeśli nie trzeciorzędnym elementem analiz. W żadnym wypadku nie chciałabym tutaj legitymizować NHST, p-wartości lub testów istotności (jestem daleka od tego), jednakże równie daleka jestem od kategorycznych dyrektyw (przynajmniej na tym etapie rozwoju badacza).

NHST oraz testy istotności utarły się w praktyce badawczej, więc trudno o to, aby jednym artykułem odczarować rzeczywistość i tok rozumowania wszystkich psychologów. Zmiany w sposobie myślenia są procesem długofalowym – nawet dzisiaj oczywiste fakty z fizyki musiały poczekać, aż społeczeństwo zaaprobuje je, a ich autorzy życie jeśli nie postradać, to mieć je uczynione nieznośnym (mam na myśli tutaj Galileusza i jego heliocentryczne poglądy). Z jednej strony mamy zatem dobro nauki, a z drugiej – właściwości psychiki człowieka. Jak zatem postąpić? Rewolucje rzadko kiedy bywają bezkrwawe, a jedynym wyjściem jest dyskusja i metoda małych kroczków.

Stąd, w moim odczuciu, dobrym, w sensie mniej szokowym, rozwiązaniem jest odejście przynajmniej od części błędów obecnych w NHST i testach istotności. Pod tym względem obiecujący staje się bootstrap oraz testowanie hipotez metodą bootstrapu. Dzięki niemu możemy zrezygnować z założeń dotyczących rozkładu statystyki testowej (np. w teście *t*-Studenta statystyka testowa ma rozkład *t*-Studenta z ustaloną liczbą stopni swobody). Bootstrap pozwala zarzucić twierdzenia działające dla prób o nieskończonej liczbie elementów, którymi wspomagamy się w analizach. Mowa tutaj m.in. o centralnym twierdzeniu granicznym.

Testowanie hipotez metodą bootstrapu jest atrakcyjne, choć dzieli z testami istotności pewne ograniczenia. Jest to raczej *tratwa* – sposób na w miarę łagodną przesiadkę z tonącego Titanica (NHST, p-wartość, testy istotności) na supernowoczesny i zaawansowany technologicznie okręt, ponieważ to właśnie statystyka bayesowska oferuje odpowiedź na stawiane przez badacza pytanie o prawdopodobieństwo prawdziwości hipotezy zerowej przy uzyskanych danych.

LITERATURA CYTOWANA

American Psychological Association. (2011). *The publication manual of the American Psychological Association* (wyd. 6). Washington: American Psychological Association.

Bańko, M. (2000). *Inny słownik języka polskiego*, t. I: A–Ó. Warszawa: Wydawnictwo Naukowe PWN.

Berger, J. O. (2003). Could Fisher, Jeffreys and Neyman have agreed on testing? *Statistical Science*, *18* (1), 1–32.

Cohen, J. (1992). Power primer. *Psychological Bulletin*, *112*, 155–159.

Cohen, J. (1994). The earth is round (p < 05). *American Psychologist*, *49*, 997–1003.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*, 7–29, doi: 10.1177/0956797613504966.

Dienes, Z. (2008). *Understanding psychology as a science: An introduction to scientific and statistical inference*. Michigan: Palgrave MacMillan.

Dienes, Z. (2016). How Bayes factors change scientific practice. *Journal of Mathematical Psychology*, *72*, 78–89.

Ellis, P. D. (2010). *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results.* Cambridge: Cambridge University Press.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London*, *222*, 309–368, doi: 10.1098/rsta.1922.0009.

Fisher, R. A. (1955). Statistical methods and scientific induction. *Journal of the Royal Statistical Society (B)*, *17*, 69–77.

Fisher, R. A. (1959). *Statistical and scientific inference* (wyd. 2). New York: Hafner Publishing Company.

Fisher, R. A. (1971). *The design of experiments.* New York: Hafner Publishing Company.

Gelman, G. (2012). P values and statistical practice. *Epidemiology*, *24*, 69–72.

Gigerenzer, G. (1993). The Superego, the Ego, and the Id in statistical reasoning. W: G. Keren, C. Lewis (red.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (s. 311–339). Hillsdale: Erlbaum.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, *33*, 587–606.

Halpin, P. F., Stam, H. J. (2006). Inductive inference or inductive behavior: Fisher and Neyman–Pearson approaches to statistical testing in psychological research (s. 1940–1960). *American Journal of Psychology*, *119*, 625–653.

Hubbard, R., Bayarri, M. J. (2003). Confusion over measures of evidence (p's) versus errors (α's) in classical statistical testing. *American Statistician*, *57*, 171–182.

Jarmakowski-Kostrzanowski, T., Jarmakowska-Kostrzanowska, L. (2014). Polska adaptacja kwestionariusza *Moral Foundations Questionnaire* (MFQ-PL). Poster zaprezentowany na XI Zjeździe Polskiego Stowarzyszenia Psychologii Społecznej. Warszawa.

Lenhard, J. (2006). Models and statistical inference: The controversy between Fisher and Neyman–Pearson. *British Society for the Philosophy of Science*, *57*, 69–91, doi: 10.1093/bjps/axi152.

Lew, M. (2012). Bad statistical practice in pharmacology (and other basic biomedical disciplines): You probably don't know P. *British Journal of Pharmacology*, *166*, 1559–1567.

Lew, M. J. (2013). To P or not to P: On the evidential nature of P-values and their place in scientific inference. Pobrane z: http://arxiv.org/abs/1311.0081.

Mayo, D. (1992). Did Pearson reject the Neyman–Pearson philosophy of statistics. *Synthese*, *90*, 233–262.

Neyman, J., Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. Philosophical transactions

of the Royal Society of London. *Series A, Containing Papers of a Mathematical or Physical Character*, *231*, 289–337.
Nosek, B. A., Spies, J. R., Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science, 7*, 615–631, doi: 10.1177/1745691612459058.
Nuzzo, R. (2014). Scientific method: Statistical errors. *Nature, 506*, 150–152, doi: 10.1038/506150a.
Pearson, E. S. (1966). *The selected papers of E. S. Pearson.* Cambridge: Cambridge University Press.
Royall, R. (1997). *Statistical evidence: A likelihood paradigm*. New York: Chapman & Hall.
R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna. Pobrane z: www.R-project.org.
Trafimow, D. (2003). Hypothesis testing and theory evaluation at the boundaries: Surprising insights from Bayes's theorem. *Psychological Review*, *110* (3), 526–535.
Trafimow, D., Marks, M. (2015). Editorial. *Basic and Applied Social Psychology*, *37*, 1–2.
Wagenmakers, E-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*, 779–804.
Wainer, H. (1999). One cheer for null hypothesis significance testing. *Psychological Methods*, *6*, 212–213.
Zabell, S. L. (1992). R. A. Fisher and the aiducial Argument. *Statistical Science*, *7*, 369–387.

Załączniki

A. Prawdopodobieństwo warunkowe

$P(A|B)$ to prawdopodobieństwo warunkowe, jest to prawdopodobieństwo zdarzenia A pod warunkiem zdarzenia B. Celem jest pokazać, że prawdopodobieństwa warunkowe $P(A|B)$ oraz $P(B|A)$ nie są sobie równe na przykładzie jednokrotnego rzutu kostką. Zdarzenie A polega na wyrzuceniu parzystej liczby oczek. Są trzy zdarzenia (elementarne) sprzyjające zdarzeniu A: otrzymanie liczby oczek 2, otrzymanie 4 lub otrzymanie 6 oczek na kostce podczas jednego rzutu, stąd $P(A) = \frac{3}{6} = \frac{1}{2}$.

Zdarzenie B polega na wyrzuceniu dwóch oczek. $P(B) = \frac{1}{6}$.

Aby obliczyć prawdopodobieństwa wyrzucenia dwóch oczek, jeśli wiadomo, że wypadła parzysta liczba oczek, należy skorzystać z równania:

$$P\left(B\middle|A\right) = \frac{P\left(B \cap A\right)}{P\left(A\right)} = \frac{P\left(A \cap B\right)}{P\left(A\right)},$$

gdzie $P(A \cap B)$ oznacza prawdopodobieństwo iloczynu (części wspólnej) dwóch zbiorów A oraz B. Iloczyn zbiorów jest przemienny, dlatego też $P(A \cap B) = P(B \cap A)$. Część wspólna zbiorów A i B, $A \cap B$, to zdarzenie polegające na wyrzucenie dwóch oczek, stąd $P(A \cap B) = \frac{1}{6}$.

Dzięki równaniu otrzymać można wynik:

$$P\left(B\middle|A\right) = \frac{\frac{1}{6}}{\frac{1}{2}} = \frac{1}{6} \cdot \frac{2}{1} = \frac{1}{3}.$$

Prawdopodobieństwo (warunkowe) otrzymania dwóch oczek, o ile wiadomo, że wypadła parzysta liczba oczek, wynosi jedna trzecia.

Prawdopodobieństwo warunkowe „w drugą stronę", czyli $P(A|B)$ oznaczałoby prawdopodobieństwo wyrzucenia parzystej liczby oczek, o ile wcześniej podano informację, że wypdła „dwójka". Zdrowy rozsądek podpowiada oczywistą odpowiedź, że takie prawdopodobieństwo wynosi 1 – poniższe rachunki to uzasadniają.

$$P\left(A\middle|B\right) = \frac{P\left(A \cap B\right)}{P\left(B\right)} = \frac{\frac{1}{6}}{\frac{1}{6}} = 1 \neq \frac{1}{3}.$$

Dzięki prostemu przykładowi widać, że prawdopodobieństwa warunkowe mogą różnić się $P(A|B) \neq P(B|A)$. Zgodnie z tym, *p – wartość* = $P(dane|H_0)$ jest zupełnie innym prawdopodobieństwem niż prawdopodobieństwo prawdziwości hipotezy zerowej pod warunkiem danych $P(H_0|dane)$, które jest dla badacza ważniejszym wskaźnikiem, czy postawiona przez niego hipoteza jest słuszna.

B. Porównanie testowania hipotez za pomocą lemat Neymana-Pearsonai podejścia Fishera

Załóżmy, że mamy próbę: $X_2, \ldots, X_{30} \sim N(\mu,1)$, stąd: odchylenie standardowe wynosi $\sigma = 1$, a liczebność $N = 30$.
Hipotezy (zerowa i alternatywna) są postaci: H_0: $\mu_0 = 0$ przeciwko H_1: $\mu = 1$.
Zgodnie z lematem N-P, obliczamy najpierw obie funkcje wiarygodności:

$$L\left(\mu_0 = 0\middle|x\right) = \frac{1}{\sqrt{2\pi}\cdot 1}\cdot exp\left\{\frac{-\left(x_1-0\right)^2}{2\cdot 1^2}\right\}\cdot\frac{1}{\sqrt{2\pi}\cdot 1}\cdot exp\left\{\frac{-\left(x_2-0\right)^2}{2\cdot 1^2}\right\}\cdots\frac{1}{\sqrt{2\pi}\cdot 1}\cdot exp\left\{\frac{-\left(x_{30}-0\right)^2}{2\cdot 1^2}\right\} =$$

$$= \left(\frac{1}{\sqrt{2\cdot\pi}}\right)^{30}\cdot exp\left\{-\frac{1}{2}\cdot\sum\nolimits_{i=1}^{30}\left(x_i-0\right)^2\right\} = \left(\frac{1}{\sqrt{2\cdot\pi}}\right)^{30}\cdot exp\left\{-\frac{1}{2}\cdot\sum\nolimits_{i=1}^{30}x_i^2\right\}$$

$$L\left(\mu_1 = 1\middle|x\right) = \frac{1}{\sqrt{2\pi}\cdot 1}\cdot exp\left\{\frac{-\left(x_1-1\right)^2}{2\cdot 1^2}\right\}\cdot\frac{1}{\sqrt{2\pi}\cdot 1}\cdot exp\left\{\frac{-\left(x_2-1\right)^2}{2\cdot 1^2}\right\}\cdots\frac{1}{\sqrt{2\pi}\cdot 1}\cdot exp\left\{\frac{-\left(x_{30}-1\right)^2}{2\cdot 1^2}\right\} =$$

$$= \left(\frac{1}{\sqrt{2\cdot\pi}}\right)^{30}\cdot exp\left\{-\frac{1}{2}\cdot\sum\nolimits_{i=1}^{30}\left(x_i-1\right)^2\right\} = \left(\frac{1}{\sqrt{2\cdot\pi}}\right)^{30}\cdot exp\left\{-\frac{1}{2}\cdot\sum\nolimits_{i=1}^{30}\left(x_i-1\right)^2\right\}$$

Lemat N-P „nakazuje" podzielić obie funkcje:

$$\frac{L\left(\mu_0 = 0\middle|x\right)}{L\left(\mu_1 = 1\middle|x\right)} = \frac{\left(\frac{1}{\sqrt{2\pi}}\right)^{30}}{\left(\frac{1}{\sqrt{2\pi}}\right)^{30}}\cdot\frac{exp\left\{-\frac{1}{2}\sum\nolimits_{i=1}^{30}x_i^2\right\}}{exp\left\{-\frac{1}{2}\sum\nolimits_{i=1}^{30}\left(x_i-1\right)^2\right\}} = exp\left\{-\frac{1}{2}\sum\nolimits_{i=1}^{30}x_i^2+\frac{1}{2}\sum\nolimits_{i=1}^{30}\left(x_i-1\right)^2\right\}\le k$$

$$-\frac{1}{2}\sum\nolimits_{i=1}^{30}x_i^2+\frac{1}{2}\sum\nolimits_{i=1}^{30}\left(x_i-1\right)^2\le\log k$$

$$-\frac{1}{2}\sum\nolimits_{i=1}^{30}x_i^2+\frac{1}{2}\sum\nolimits_{i=1}^{30}\left(x_1^2-2\cdot x_i+1\right)\le\log k$$

$$-\frac{1}{2}\sum\nolimits_{i=1}^{30}x_i^2+\frac{1}{2}\sum\nolimits_{i=1}^{30}x_i^2-2\cdot\frac{1}{2}\sum\nolimits_{i=1}^{30}x_i+\frac{1}{2}\sum\nolimits_{i=1}^{30}1\le\log k$$

$$-\frac{1}{2}\sum\nolimits_{i=1}^{30}x_i^2+\frac{1}{2}\sum\nolimits_{i=1}^{30}x_i^2-\sum\nolimits_{i=1}^{30}x_i+30\le\log k$$

$$-\sum\nolimits_{i=1}^{30}x_i+30\le\log k$$

$$\frac{\sum_{i=1}^{30}x_i}{30}\ge\frac{-\log k+30}{30}$$

Zastępujemy prawą stronę nierówności jednym liczbą k^*

$$\frac{\sum_{i=1}^{30}x_i}{30}\ge k^*$$

Prawa strona równania zależy jedynie od próby i można ją uznać za statystykę testową:

$$T(X) = \frac{\sum_{i=1}^{30} x_i}{30} = \overline{x}.$$

Własności probabilistyczne suma zmiennych losowych o rozkładzie normalnym uprawniają do stwierdzenia, że suma również posiada rozkład normalny z podanymi parametrami:

$$T(X) \sim N\left(0, \frac{\sigma}{\sqrt{n}}\right).$$

Wykorzystując wiedzę o liczebności próby i wariancji, można napisać: $n = 30$, $\sigma = 1$ więc $= \frac{\sigma}{\sqrt{30}} = \frac{1}{\sqrt{30}}$.

Zapis matematyczny prawdopodobieństwo popełnienia błędu I rodzaju, alfa, pomaga w znalezieniu wartości krytycznej w zbiorze możliwych wartości statystyki testowej:

$$\alpha = P(T(X) > k^*|H_0) = 0{,}05$$

Przekształcając powyższą równość otrzymamy, zapisane matematycznie, polecenie znalezienia odpowiedniego kwantyla:

$$P(T(X) < k^*|H_0) = 0{,}95$$

Aby podać dokładną wartość krytyczną k^*, należy znaleźć kwantyl rzędu 0,95 rozkładu normalnego. Wynosi on 0,3, więc $k^* = 0{,}3$. Jeśli statystyka testowa obliczona na podstawie zebranej próby, przekroczy tę wartość, to będzie można odrzucić hipotezę zerową H_0: $\mu = 0$ i przyjąć, że próba pochodzi z rozkładu normalnego o średniej $\mu = 1$.

Przechodząc teraz do części praktycznej, przy pomocy programu R wygenerowano hipotetyczną próbę z rozkładu standardowego normalnego ~ $N(0,1)$:
0,32; –1,77; –1,66;1,42;1,20; –0,33; –0,81; –0,55; –0,01; –1,12; 0,61; 0,37; –0,35; 0,41; –0,76; –0,77; –0,92; 1,08; –0,71; –0,39; 0,82; 1,10; –0,17; –0,70; 1,34; 2,75; –0,39; –0,16; 0,42; –0,44.

Statystyka testowa, będąca średnią z próby, $T(X) = \overline{x}$ dla powyższej próby wynosi –0,005. Jest to wartość mniejsza niż 0,03. Stąd też wybrano hipoteza zerowa H_0: $\mu = 0$.

Wygenerowano również inną próbę pochodzącą z rozkładu normalnego ~ $N(1,1)$ o parametrze położenia równym 1: 2,41; 1,26; 0,82; 0,88; 0,19;0,13; 0,08; 1,78; 0,94; 2,06; 0,96; –0,30; –0,18; 1,39; 1,65; 0,55; –0.54; –0,27; 2,09; 1,92; 0,67; 2,12; 0,03; 0,68; –0,06; 0,78; 0,73; 0,19; 0,70; –0,04.

Dla tej próby wartość statystyki testowej $T(X) = \overline{x} = 0{,}79$. Ze względu na to, że 0,79 > 0,3, można podjąć decyzję o odrzuceniu hipotezy zerowej i przyjąć hipotezę alternatywną, H_1: $\mu = 1$. Próba pochodzi z rozkładu normalnego o średniej równej 1.

Z kolei w podejściu Fishera obliczana jest najpierw statystyka testowa $T(X) = \frac{\overline{x} - 0}{\frac{1}{\sqrt{30}}}$ (nieco inna niż w lemacie N-P). Następnie, przy założeniu prawdziwości hipotezy zerowej rozkład tej statystyki jest znany (rozkład *t*-Studenta z 29 stopniami swobody). Stąd można wykonać zadanie obliczenia p-wartości, czyli prawdopodobieństwo znalezienia wyniku takiego jak otrzymany lub bardziej ekstremalnego ($P(T(X) > \overline{x}|\mu = 0)$) – odczytać je z tablic rozkładu t-Studenta, albo sprawdzić w programie statystycznym.

Dla pierwszej próby, p-wartość wynosi 0,98, stąd nie ma podstaw do odrzucenia hipotezy zerowej. Wartość oczekiwana rozkładu normalnego cechy w badanej populacji wynosi 0.

W drugiej próbie, p-wartość jest równa $1{,}538 \cdot 10^{-0{,}05} = 0{,}000015$, co powoduje odrzucenie hipotezy zerowej.

In the statistical matrix: Null hypothesis significance testing and p-value controversies

Lilianna Jarmakowska-Kostrzanowska

Interdisciplinary Center of Modern Technology, Nicolaus Copernicus University in Toruń

Abstract

Statistical analyses in a vast majority of psychological research are based on null hypothesis significance testing (NHST) – a several-step procedure which is considered an invalid mixture of two incompatible approaches – one of Fisher, the other of Neyman–Pearson. Despite that many advocates abandoning NHST, a lot of researchers, psychologists among them, still use this flawed routine. The article presents ideas, which NHST is based on, and also controversies around its most famous construct, p-value. It also discusses objections against p-value (e.g., dependence on sample size) and ways of interpreting it (e.g., conditional probability). The main task of the article is to present the basic concepts of NHST in an understandable manner for Polish social science researchers.

Keywords: *null hypothesis significance testing (NHST)*, *p-value*, *hypothesis testing*

Złożono: 8.05.2015
Złożono poprawiony tekst: 12.10.2015/7.12.2015
Zaakceptowano do druku: 25.01.2016