PRZEGLĄD PSYCHOLOGICZNY, 2021, TOM 64, Nr 4, 83–99
DOI: 10.31648/pp.7359

Istotnie statystyczna moc testu – analiza mocy i jej miejsce w przyborniku badacza oraz interpretacja (nie)istotności statystycznej przy małej (dużej) mocy testu

Lilianna Jarmakowska-Kostrzanowska*
Instytut Psychologii
Uniwersytet Mikołaja Kopernika w Toruniu
0000-0003-3644-006X

STRESZCZENIE

Cel
W prezentowanym artykule przyjęto dwa zasadnicze cele – zarysowanie problematyki mocy testu statystycznego oraz omówienie podstawowych problemów związanych z analizą mocy nowym–starym narzędziem. Nowym – w typowym warsztacie badacza, lecz jednocześnie starym, ponieważ znanym już w statystyce. W pracy omówiono także techniczną stronę analizy mocy i jej umiejscowienia względem p-wartości.

Tezy
Moc i analiza mocy oraz istotność statystyczna są pojęciami z dwóch różnych szkół statystyki, które łącznie tworzą paradygmat testowania istotności hipotezy zerowej (NHST). Niespójność szkół składających się na ten paradygmat przysparza problemów w interpretacji wyników testu.

Konkluzje
Wprawdzie analiza mocy pozwala na wyznaczenie potrzebnej wielkości próby, ale późniejsza interpretacja wyników testu nie jest łatwa. Trudno jednoznacznie wskazać, jak należy interpretować wynik nieistotny statystycznie przy dużej mocy testu lub istotny statystycznie wynik przy małej mocy testu. Oprócz tego moc testu ani nie uprawdopodabnia wyniku istotnego statystycznie wyniku, ani też nie obala hipotezy zerowej, gdy wynik jest nieistotny statystycznie.

Słowa kluczowe: istotność statystyczna, p-wartość, analiza mocy, moc testu

* Adres e-do kontaktu: Lilianna Jarmakowska-Kostrzanowska, Instytut Psychologii UMK, Uniwersytet Mikołaja Kopernika w Toruniu, ul. Jurija Gagarina 39, 87-100 Toruń; e-mail: lkostrzanowska@umk.pl.

W badaniach ilościowych kluczowe pytanie o liczbę obserwacji (wielkość próby *sample size*) jest traktowane jako typowe, wręcz zwyczajne. Mimo to przez długi czas zagadnienie dotyczące liczby osób badanych znajdowało się poza spektrum zainteresowań badacza. Próby liczyły tyle osób, ile można było ich pozyskać. Ta swoboda trwała do czasu, gdy w psychologii rozwinął się kryzys replikacyjny (Ioannidis, 2005), który postawił pod znakiem zapytania opisywane w literaturze zjawiska.

Nauka opiera się na replikacjach. Jeśli rezultatów oryginalnego badania nie udaje się powtórzyć, można przypuszczać, że było ono artefaktem lub przeszacowanym wynikiem. Tak stało się w przypadku hipotezy mimicznego sprzężenia zwrotnego (*facial feedback hypothesis*; Strack, Martin, Stepper, 1988). Hipoteza ta głosi, że trzymając w ustach długopis, rozciągamy mięsień jarzmowy, a to powoduje subiektywny wzrost poczucia radości. Hipotezę tę sformułowano na podstawie badań, których wyników nie udało się powtórzyć (Wagenmakers i in. 2016). Podobnie rzecz się ma w przypadku badania nad aktywacją myślenia analitycznego dotyczącego czytania tekstu napisanego czcionką utrudniającą jego płynne czytanie. Pierwotne badanie pokazywało, że jeśli studentom prezentowano materiał z czcionką utrudniającą czytanie, wówczas popełniali oni mniej błędów. Również i tego eksperymentu nie udało się powtórzyć (Meyer i in., 2015; Sirota i in., 2020). Nie są to jedyne badania, których oryginalne wyniki okazały się bardziej obiecujące niż ich replikacje (zob. Klein i in., 2014).

Za jeden z powodów kryzysu replikacyjnego uważa się brak dbałości o wielkość próby. Natomiast za jedno z rozwiązań – analizę mocy. Wprawdzie to pojęcie znano już wcześniej (np. Cumming, 2011; Murphy, Myors i Wolach, 2014), ale dopiero ostatnio zdobywa ono popularność, generując jednak nowe problemy i nowe wyzwania. Jednocześnie w przyborniku pozostaje istotność statystyczna. Istotność statystyczna oraz poziom mocy testu są i będą jednocześnie wykorzystywane do interpretacji otrzymanego wyniku. Przedmiotem tego artykułu są kwestie związane z mocą testu, a w szczególności, czym ona jest, jak działa, a także to, co dzieje się po wykonaniu analiz, czyli interpretacja wyników istotnych bądź nieistotnych statystycznie przy niskiej lub dużej mocy testu.

DEFINICJE MOCY STATYSTYCZNEJ

Problem z nowym narzędziem zaczyna się w miejscu, gdzie, co do zasady, powinna panować jednoznaczność w definicji mocy testu. Literatura naukowa dysponuje określeniami tego pojęcia. Według niektórych autorów moc to prawdopodobieństwo uniknięcia błędów II-ego rodzaju (Neyman, 1977). Inni, szczególnie w badaniach medycznych, wskazują, że moc to zdolność do uniknięcia fałszywie ujemnego wyniku (*false negative*). Z kolei Cohen (1988, s. 1) określa ją jako prawdopodobieństwo uzyskania istotnego statystycznie wyniku.

Pierwsza definicja została oparta na innych pojęciach stworzonych przez Neymana i Pearsona, co powoduje, że osoby nieznające ich podejścia ocierają się o *ignotum per ignotum* (nieznane przez nieznane). Druga definicja wprawdzie

zrównuje test statystyczny z testem medycznym, ale pozwala intuicyjnie zarysować pojęcie osobie niezwiązanej ze statystyką. Przystępnie wyrażona trzecia definicja zawiera jednak trudną do usunięcia wadę (Mayo, 2018, s. 324). Przyswojenie właśnie tej definicji utrudnia prawidłową interpretację wyników badania. Aby dostrzec tę wadę i jej konsekwencje, należy przyjrzeć się testowaniu hipotez statystycznych w psychologii, czemu poświęcono następną sekcję artykułu. Przegląd pojęć pozwoli czytelnikowi zapoznać się z miejscem analizy mocy w zbiorze dotychczasowych technik statystycznych.

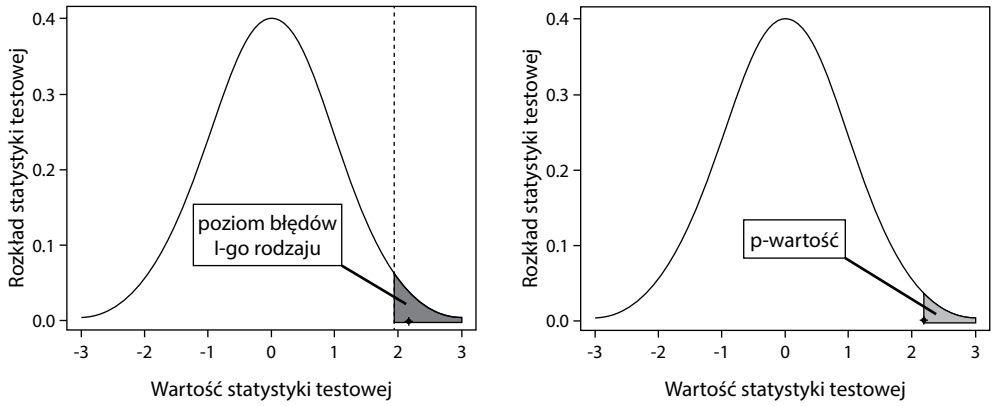
Moc statystyczna i istotność statystyczna

Zacznijmy od prostego spostrzeżenia, że badacz nauk empirycznych musi liczyć się ze zmiennością wyników swojego eksperymentu. Oznacza to, że nawet poprawnie zbudowany model wiążący zmienne nie pozwala na napisanie deterministycznych wzorów znanych z chemii czy fizyki w postaci y = f(x). Psycholog nie może powiedzieć, że każda frustracja kończy się agresją u każdego człowieka.

Aby przeanalizować wyniki, psychologowie sięgają po procedury testowania istotności statystycznej hipotezy zerowej (*Null hypothesis significance testing*, w skrócie: NHST; patrz: Wolski, 2017, Jarmakowska-Kostrzanowska, 2016; Gigerenzer, 2004). Gdyby prześwietlić NHST, okazałoby się, że złożono ją z dwóch osobnych szkół: frekwentystycznej i fisherowskiej, a każda z nich wypracowała własne pojęcia i przede wszystkim – sposób interpretacji wyników. Mimo wielu różnic przedstawiciele dwóch szkół zgadzają się co do jednego: hipotez nie weryfikuje się przez bezpośredni ogląd danych w arkuszu kalkulacyjnym, a przez ich przekształcenie, zwane statystyką testową. Ponieważ dane są losowe, więc i wyniki statystyki testowej są losowe. Do dalszego wnioskowania należy poznać rozkład szans pojawiania się wyników. To z kolei wymaga przyjęcia pewnych założeń odnośnie do hipotez. Jeśli założy się prawdziwość hipotezy zerowej o braku związku między zmiennymi, to rozkład statystyki testowej jest znany. Przykładowo test t-Studenta ma *n-1* stopni swobody przy założeniu braku różnic między grupami (*n* to liczba obserwacji).

W dalszej części tekstu w celu uproszczenia będziemy posługiwać się jednopróbowym testem t-Studenta *one sample t-test*. Typowa sytuacja analizowana za jego pomocą to taka, w której badacz sprawdza, czy uzyskana średnia w próbie zgadza się z teoretyczną wartością, np. czy ciśnienie tętnicze jest równe książkowemu poziomowi 120 Hg.

Oba panele rysunku 1 (s. 86) przedstawiają krzywą rozkładu statystyki testowej t-Studenta dla n = 50 osób badanych, gdy hipoteza zerowa jest prawdziwa. To, co różni te rysunki, to sposób użycia krzywych w każdej ze szkół. Lewa strona ilustruje wykorzystanie rozkładu statystyki testowej t-Studenta, gdy badacz pracuje według założeń szkoły neymanowskiej, a prawa, gdy odnosi się do szkoły fisherowskiej. Według pierwszej, w której hipotezę alternatywną przeciwstawia się hipotezie zerowej, badacz ustala odgórnie odsetek błędów polegających na uznaniu, że odkrył zjawisko, choć w rzeczywistości jest to artefakt. My znamy to pod nazwą błędu I-ego rodzaju, czyli błędnego odrzucenia prawdziwej hipotezy zerowej o braku



Rysunek 1. Statystyka testowa i jej rozkład wobec dwóch szkół myśli statystycznej.

Nota. Lewa strona obrazuje podejście frekwentystyczne: Ustalony przed badaniem poziom błędu I-go rodzaju wyznacza region odrzuceń statystyki testowej (czarna kropka), niezależnie od tego, w którym punkcie znalazła się. Prawa strona ilustruje podejście fisherowskie: Punkt położenia wartości statystyki (czarna kropka) jest ważny, ponieważ stanowi podstawę obliczeń p-wartości. Źródło: opracowanie własne.

zależności. Zwyczajowo badacz żąda, aby mylić się nie więcej niż 5% razy, stąd przerywaną linią zaznaczona jest krytyczna wartość statystyki testowej. W przypadku $n = 50$ obserwacji wynosi ona po zaokrągleniu 1,68. Wyższe od niej wartości powodują odrzucenie hipotezy zerowej na korzyść alternatywnej. Uzyskana przez badacza wartość statystyki testowej (symbolizowana czarną kropką) sama w sobie nie ma znaczenia – ważne jest to, czy znalazła się w szarym polu. Dlatego podejście Neymana nazywa się *fixed-alpha approach* (Huberty, 1993). W szkole fisherowskiej po prawej stronie brakuje ustalonego poziomu błędu I-ego rodzaju, ponieważ sama szkoła nie ma tego konceptu. Tutaj wartość statystyki testowej, zaznaczona czarną kropką, ma ogromne znaczenie. To na jej podstawie jest obliczana p-wartość, czyli pole na prawo od czarnej kropki. Jeśli obszar pola jest mniejszy niż 5% całości, to wynik statystyki testowej jest istotny statystycznie. Dlatego podejście Fishera nazywa się *p-value approach* (Huberty, 1993).

Zauważmy, że do momentu obliczenia wartości statystyki testowej badacz nie zwraca uwagi na to, w obrębie jakiej szkoły pracuje. Z punktu widzenia użytkownika statystyki podejścia są nie do odróżnienia: te same statystyki testowe, podobne nazewnictwo i referencyjne wartości narzędzi[1].

Nakładając wykresy jeden na drugi, staje się jasne, dlaczego badacze skłaniają się do interpretacji fisherowskiego poziomu istotności statystycznej jako neymanowskiego błędu I-ego rodzaju, o czym piszą Hubbard i Bayarri (Hubbard i Bayarri, 2003). Pozostaje pytanie, czy można podobny błąd popełnić w drugą

[1] Chodzi o fisherowski poziom istotności i neymanowski błąd I-go rodzaju – w obu przypadkach wynosi on 5%.

stronę, tym razem połączyć istotność statystyczną z błędem II-ego rodzaju, jak w definicji Cohena?

Wróćmy do lewego panelu rysunku 1. Jednym z ważnych elementów szkoły Neymana-Pearsona jest błąd II-ego rodzaju (a co za tym idzie: moc). Błąd II-ego rodzaju to błąd badacza, który błędnie przyjmuje, że w danych nic nie ma. Dokonuje wówczas błędnego odrzucenia prawdziwej hipotezy alternatywnej o istnieniu zależności. Z tym kluczowym dla szkoły neymanowskiej konceptem nie koresponduje żaden element rysunku 1. Jak zatem obliczyć błąd II-ego rodzaju? Wkrótce okaże się, że na rysunku brakuje jeszcze jednej krzywej reprezentującej rozkład szans statystyki testowej – odpowiadającej prawdziwej hipotezie alternatywnej.

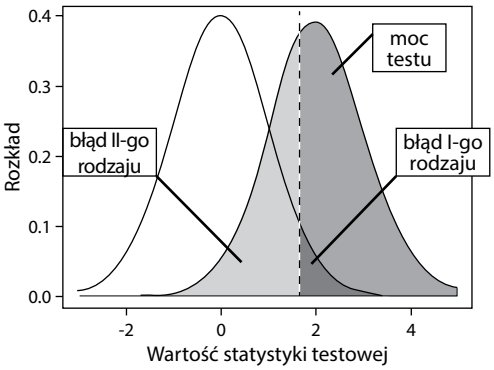
Każda statystyka testowa ma dwa rozkłady: jeden występuje w warunkach prawdziwości hipotezy zerowej, drugi – w warunkach prawdziwości hipotezy alternatywnej. Narysowanie krzywej dla tego drugiego rozkładu wymaga wprowadzenia dodatkowego parametru do rozkładu statystyki testowej, zwanego parametrem niecentralności *noncentrality parameter*. Badacz pracujący tylko w paradygmacie NHST nie ma okazji zetknąć się ani z tym pojęciem, ani z błędem II-ego rodzaju, ani z mocą testu. Krzywe rozkładów statystyki testowej dla hipotezy zerowej nie potrzebują podawania parametru niecentralności.

Parametr niecentralności to liczba determinująca kształt całej krzywej (Kelley, 2013), zależna od innych wielkości, np. liczby obserwacji i tego, jaką wartość postuluje hipoteza alternatywna. O ile liczba obserwacji musi być pojedynczą liczbą, o tyle zwyczajowo stawiana hipoteza alternatywna w ramach NHST jest niespecyficzna (np. H_1: $\mu \neq 0$ lub H_1: $\mu > \mu_1$ lub H_1: $\mu_1 \neq \mu_2$). Trudno wówczas oczekiwać, że pojedyncza krzywa będzie odpowiadać całemu spektrum wartości spełniającym nierówność zawartą w hipotezie alternatywnej. Zatem należy podać konkretną postać proponowanego parametru, np. H_1: $\mu = 0{,}2$ lub H_1: $\mu_1 - \mu_2 = 0{,}5$. Parametr niecentralności stanie się pojedynczą liczbą i dopiero wówczas program dorysuje drugą krzywą – zwaną krzywą rozkładu statystyki testowej dla prawdziwej hipotezy alternatywnej. Nadanie konkretnej, w matematycznej terminologii, punktowej wartości poszukiwanego parametru ma poważne konsekwencje podczas przeprowadzenia analizy mocy. Badacz stanie przed zadaniem wyboru odpowiedniej wartości – jednej liczby z całego zakresu możliwości.

W naszym przykładzie ustalmy, że H_1: $\mu = 0{,}5$ (to arbitralnie wybrana wartość) i spójrzmy na rysunek poniżej.

Rysunek 2 (s. 88) jest typowym graficznym przedstawieniem zależności między pojęciami statystyki klasycznej: błędami I i II rodzaju oraz zależnością między mocą a błędem II-go rodzaju. Błąd I-go α i II-go rodzaju β nie wiąże żadna prosta arytmetyczna zależność. Oba pola liczone są na podstawie różnych krzywych i nie sumują się do żadnej liczby. Ponieważ znajdują się po obu stronach krytycznej wartości statystyki testowej, to następuje ich sprzężenie. Im większy błąd beta, tym mniejszy błąd alfa i na odwrót – im większy błąd alfa, tym mniejszy błąd beta.

Prosta arytmetyczna zależność istnieje między mocą a błędem II-go rodzaju, ponieważ zajmuje obszar znajdujący się pod tą samą krzywą. Ta zależność to moc = 1 – beta. Im większe pole odpowiadające błędowi II-go rodzaju, tym mniejsza moc takiego testu.



Rysunek 2. Wzajemne relacje błędu I-go rodzaju, alfa oraz błędu II-go rodzaju, beta

Nota. Rozkłady statystyki testowej dla prawdziwej hipotezy zerowej (pierwsza od lewej) i alternatywnej (gdy H_1: $\mu = 0{,}5$; pierwsza od prawej)). Źródło: opracowanie własne.

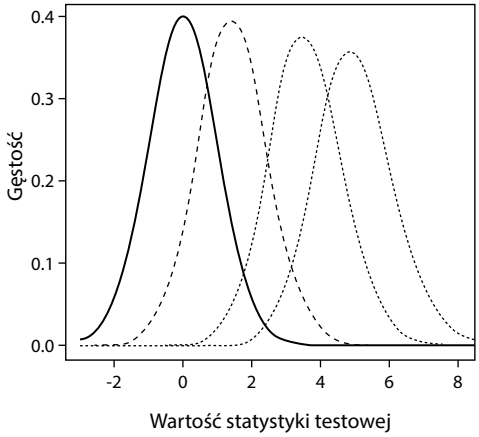
W tym miejscu należy wyjaśnić, dlaczego ponowna kolizja słowników szkół statystycznych widoczna w definicji Cohena jest technicznie możliwa. Zwróćmy uwagę na prawą stronę kreskowanej linii. Wzdłuż osi poziomej leżą wartości statystyki testowej, które w terminologii Neymana powodują odrzucenie hipotezy zerowej, a dla Fishera są wynikami istotnymi statystycznie. Znajdująca się nad nimi krzywa zakreśla pole odpowiadające szansom wystąpienia wyniku z tego zakresu. Tym samym neymanowska moc staje się prawdopodobieństwem fisherowskiego wyniku istotnego statystycznie – właśnie tak definiuje ją Cohen (1988, s. 1). Nałożenie rysunków ilustrujących pojęcia z dwóch szkół pokazuje, jak łatwo połączyć te szkoły.

Techniczne aspekty można byłoby pominąć – Neyman również posługiwał się terminem istotność, choć czynił to we własnym rozumieniu. Za to o wiele ważniejszą kwestią jest zrozumienie przynależności pojęć mocy oraz istotności statystycznej do szkół statystycznych. Moc testu jest pojęciem ze szkoły neymanowskiej, frekwentystycznej. W tej szkole zakłada się, że – o ile zjawisko w rzeczywistości istnieje – dobrą decyzję podejmuje się w dużym odsetku eksperymentów, nie w pojedynczym wykonaniu badania. Fisherowska p-wartość nie ma perspektywy dużej liczby powtórzeń tego samego eksperymentu. Tu jest jedno badanie, czyli to, które wykonano i na podstawie którego badacz chce wiedzieć, co ma myśleć o hipotezie zerowej. Wydawałoby się, że spory o filozofię można pozostawić filozofom, ale właśnie owe filozoficzne podstawy generują późniejszy problem przy interpretacji, np. czy istotny statystycznie wynik przy dużej mocy jest potwierdzeniem hipotezy alternatywnej.

Warunkowość pojęcia mocy

Definicja Cohena przytoczona na początku artykułu zawiera jeszcze jeden defekt. Zauważmy, że na rysunku 1 (s. 86) bez jednoznacznie postawionej hipotezy alternatywnej nie można narysować krzywej odpowiadającej błędom II-go

rodzaju. W przykładzie przyjęliśmy, że H_1: μ = 0,5. Gdyby przyjąć inne wartości np. H_1: μ = 0,2 lub H_1: μ = 0,7, program dorysowałby krzywą innego kształtu (zob. rysunek 3).



Rysunek 3. Zmieniające się krzywe rozkładu statystyki testowej dla prawdziwej hipotezy alternatywnej o różnych parametrach położenia

Nota. Zwiększająca się wartość parametru μ działa na rozkład statystyki testowej dla prawdziwej hipotezy alternatywnej. Rozkład zmienia nie tylko kształt, ale również położenie. Źródło: opracowanie własne.

Wróćmy do rysunku 2. Przesuwając kreskowaną linię w lewo lub w prawo, a więc także zwiększając lub zmniejszając błąd I-go rodzaju, badacz steruje mocą testu. To sugeruje, że na moc ma również wpływ wielkość próby. Zatem test statystyczny sam w sobie nie ma fabrycznie wbudowanej mocy. Może mieć 30% albo 99% mocy. Wszystko zależy od tego, ile razy badacz chciałby się mylić, odrzucając prawdziwą hipotezę zerową (poziom błędów I-go rodzaju) i odrzucając prawdziwą hipotezę alternatywną (poziom błędów II-go rodzaju), a także od tego, ile osób przebadał i jaka jest wartość poszukiwanego parametru proponowana w hipotezie alternatywnej. I tak 80% moc testu występuje w dokładnie określonych okolicznościach. Tej warunkowości w przystępnej definicji mocy Cohena brakuje.

Analiza mocy – istota analizy

Liczba obserwacji, błąd I-go oraz II-go rodzaju, a także wartość postulowana przez hipotezę alternatywną stanowią układ czterech równań z czterema niewiadomymi. Ustalenie trzech z nich pozwala wyznaczyć czwartą, co wykorzystuje analiza mocy. Rachunki wykonuje darmowy program G*Power lub SPSS od wersji 27.

Wróćmy do rysunku numer 2. Po lewej stronie kreskowanej linii leży część masy rozkładu (pola pod jego krzywą) odpowiadająca za błąd II-go rodzaju.

Badaczowi zależy, aby była ona jak najmniejsza. Aby przesunąć ją na prawo, należy sterować tym, co wpływa na parametr niecentralności – liczbą obserwacji oraz poziomem błędu I-go rodzaju lub wartością parametru proponowaną w hipotezie alternatywnej.

Wielkość efektu

Ostatnim pojęciem wymagającym omówienia jest wielkość efektu (*effect size*). W analizie mocy zamiast bezpośredniej wartości parametru używa się tzw. wielkości efektu. Wielkość efektu ma rozbudowaną naukową definicję m.in. Kelleya i Preachera (Kelley i Preacher, 2012). Na potrzeby artykułu wystarczy następująca terminologia – niech *efekt* będzie synonimem zjawiska czy zależności, zaś wielkość efektu to miara siły tego zjawiska.

Wyraźnie sformułowana hipoteza alternatywna to hipoteza przyporządkowująca konkretną wartość do spodziewanej wielkości efektu, np. standaryzowana różnica między dwiema średnimi *d* Cohena wynosi 0,5 lub współczynnik korelacji *r* Pearsona wynosi 0,7. Dobór odpowiedniej wartości wielkości efektu, czyli pojedynczej liczby, spoczywa na badaczu. Jej wyznaczenie jest niełatwe. Istnieje cały zestaw sposobów doboru odpowiedniej wielkości. Badacz dokonuje więc czegoś, co w języku angielskim nazywa się domysłem opartym na posiadanej wiedzy *educated guess* – mianowicie zgaduje, jaka powinna być wartość efektu w populacji, szacując ją na podstawie dostępnych źródeł. Ma to duże znaczenie dla interpretacji wyników, ponieważ po wykonaniu badań badacz ma dwie wielkości efektu – teoretyczną, którą wskazał w analizie mocy, oraz empiryczną, którą otrzymał w badaniu.

Zwyczajowe progi błędów I-go i II-go rodzaju w analizie mocy

Analiza mocy służy do wyznaczenia potrzebnej wielkości próby, n, ale w tym celu należy określić trzy pozostałe wartości: wielkość efektu, z którą związane są omówione wyżej trudności; błąd I-go rodzaju oraz błąd II-go rodzaju. Błąd I-go rodzaju domyślnie wynosi 5%. W przypadku mocy w psychologii stosuje się referencyjną wartość Cohena równą 80% dla mocy (Cohen, 1988). Próg ten oznacza, że jeśli dany efekt istnieje, błąd II-go rodzaju wyniesie 20%. Innymi słowy, badacz myli się w 20% przypadków. W jednym na pięć odrzuca prawdziwą hipotezę alternatywną o istnieniu efektu i mylnie przyjmuje, że efekt jest zerowy.

Badania niedomocowane *underpowered studies* i o zbyt dużej mocy *overpowered studies*

Przyjęcie pojedynczej liczby jako wartości referencyjnej naturalnie dzieli badania na takie, w których użyty test ma moc mniejszą niż 80% i na większą niż 80%. Pierwszy rodzaj badań nosi nazwę badań niedomocowanych. Badanie

niedomocowane to takie, w którym użyto testu o mocy zbyt małej do wykrycia określonej wielkości efektu. Taki testu można porównać do mikroskopu o ostrości dostosowanej do dużych obiektów – nie wykryje on milimetrowych nicieni, a piętnastocentymetrową dżdżownicę już tak.

Po przyjęciu 80% za próg odpowiedniej mocy naturalne staje się pytanie: jak bardzo poniżej tego progu można zejść? Test, który ma 78% mocy, jest prawie tak samo dobry, jak ten, którego moc wynosi 80%[2]. W obu przypadkach mamy stosunkowo dużą szansę wykrycia zjawiska – o ile ono istnieje. Natomiast test, którego moc wynosi 50%, jest równie dobry jak rzut monetą. Wokół wartości 80% trudno zakreślić granice przedziału takich wartości mocy, które byłyby możliwe do przyjęcia.

Testy o niepokojąco niskiej mocy to takie, których moc spada dużo poniżej 80-procentowego progu. Po drugiej stronie znajdują się badania, które środowisko badaczy nazwało badaniami o zbyt dużej mocy. Takim terminem nazywane są badania, w których użyty test ma moc znacznie większą niż 80%.

Skoro badania o mocy zbyt małej do wykrycia zjawiska spowodowały kryzys replikacyjny, można postawić pytanie o to, jakie niebezpieczeństwo/ryzyko niosą za sobą badania, w których wykorzystano testy o mocy znacznie przekraczającej próg 80%. Tak dużą moc często osiąga się przez dużą wielkość próby, a przecież nie wszystkie badania polegają na wypełnianiu kwestionariuszy online.

Badania mogą narażać na uszczerbek na zdrowiu osoby badane, co jest kosztem natury etycznej, mogą też wymagać dużych nakładów finansowych lub wiązać się z trudnościami logistycznymi (przewóz sprzętu).

Jeśli jest możliwość ograniczenia kosztów – mamy argument przemawiający za utrzymywaniem ustalonego poziomu mocy testu (równym 80%) i nieprzekraczaniem tej granicy. Jednak zbyt duża moc nie jest problemem natury statystycznej.

INTERPRETACJA WYNIKÓW: ISTOTNOŚĆ STATYSTYCZNA I MOC STATYSTYCZNA

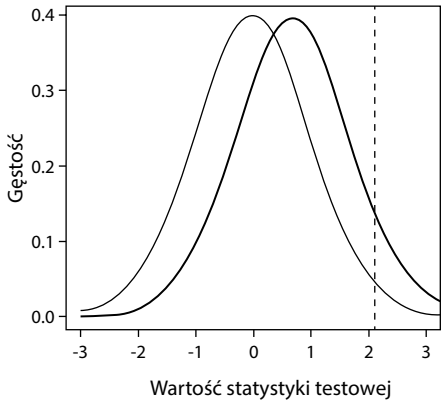
Przez długi czas istotność statystyczna była jedynym narzędziem podejmowania decyzji o prawdziwości bądź nieprawdziwości hipotezy zerowej. Typowy przybornik badacza poszerzył się o nowe narzędzie – analizę mocy. Czy dodatkowa informacja o wysokości poziomu mocy jest brakującym elementem, który pozwoli wnioskować o prawdziwości hipotezy zerowej lub alternatywnej? Badacz mógłby pomyśleć, że jeśli zaplanował 80-procentową moc testu, to otrzymanie istotnego statystycznie wyniku oznaczałoby, że jest bardzo prawdopodobne, iż jego hipoteza zerowa jest nieprawdziwa. Albo w drugą stronę – otrzymanie nieistotnego statystycznie wyniku przy dużej mocy oznaczałoby, że hipoteza zerowa jest prawdziwa. W tym miejscu można przeanalizować cztery scenariusze.

[2] Parafrazując słowa Cohena (Cohen, 1990), Bóg kocha 81% tak samo jak 79%.

Istotność statystyczna oraz niska moc testu

Zdrowy rozsądek podpowiada, że testy o małej mocy, powinny wskazywać jedynie wyniki nieistotne statystycznie (lub równoważnie: wartość statystyki testowej powinna znajdować się w zbiorze przyjęć hipotezy zerowej) – nie mają mocy wykrywania zjawiska, są zbyt słabe. Tymczasem, raz na jakiś czas, wartość statystyki testowej wpada w zbiór odrzuceń hipotezy zerowej, a p-wartość przekracza próg 0,05. Łatwo zatem skonstatować, że mimo problemów z mocą, test jednak wykrył jakiś efekt. Jak interpretować istotny statystycznie rezultat przy niskiej mocy? To pytanie sprowadza się do innego: co powoduje odrzucenie hipotezy zerowej, gdy test ma niską moc?

Można posłużyć się przykładem. Aby test t-Studenta miał moc 80% dla wielkości efektu d Cohena równej 0,1, liczba osób badanych musi wynosić ponad $n = 787$. Przypuśćmy, że ze względów czasowych badacz może przebadać jedynie $n = 50$ osób, więc moc testu spada do 10% (rysunek 4).



Rysunek 4. Rozkłady statystyk testowych dla prawdziwej hipotezy zerowej i dla prawdziwej hipotezy alternatywnej, gdy test jest małej mocy.

Nota. Cieńszą linią zaznaczono rozkład statystyki testowej, gdy hipoteza zerowa o braku efektu jest prawdziwa. Obok znajduje się grubszą linią zaznaczony rozkład statystyki testowej, gdy hipoteza alternatywna o wielkości efektu równej 0,1 jest prawdziwa. Przerywaną linią zaznaczono wartość statystyki testowej, która odpowiada za początek zbioru krytycznych wartości. Źródło: opracowanie własne.

Rysunek przedstawia krzywą rozkładu statystyki testowej, gdy hipoteza zerowa jest prawdziwa (cieńsza linia). Z paragrafu o powstawaniu rozkładów wynika, że w sytuacji, gdy hipoteza alternatywna jest prawdziwa, rozkładem steruje parametr niecentralności. Z kolei, wiadomo o nim, że koresponduje z wielkością efektu i wielkością próby. Tutaj wielkość efektu wynosi 0,1, a próba jest mała, $n = 50$. W związku z tym cała krzywa rozkładu szans dla możliwych wyników statystyk testowych, gdy hipoteza alternatywna jest prawdziwa, leży całkiem blisko (grubsza linia) krzywej odpowiadającej prawdziwej hipotezie alternatywnej.

Przerywana linia to linia oznaczająca krytyczną wartość statystyki testowej, która w takim układzie wynosi 2,10 (odczytane z tablic rozkładu statystyki testowej). Po jej lewej stronie znajduje się większość pola zakreślonego przez obie krzywe. Oznacza to, że badacz najczęściej będzie otrzymywać wartości statystyki testowej mniejsze niż 2,10. Trudno jednak będzie zdecydować, którą krzywą należy wybrać, właśnie ze względu na ich bliskie położenie. Wartości typowe dla hipotezy zerowej o braku efektu są prawie tak samo prawdopodobne, jak wartości typowe dla hipotezy alternatywnej. Tymczasem, raz na jakiś czas, znajdzie się taka wartość, która przekroczy przerywaną linię i znajdzie się po prawej stronie przerywanej linii. W jednym i w drugim przypadku (prawdziwa hipoteza zerowa lub alternatywna) będą to mniej typowe wartości, bo znajdują się w dalszych częściach obu rozkładów. W obu przypadkach potrzeba dużej wartości statystyki testowej, aby znaleźć się właśnie w tym miejscu.

A co może spowodować dużą wartość statystyki testowej? Duża statystyka testowa jest równoznaczna z dużą różnicą między średnimi oraz mniejszą zmiennością wyników. Tymczasem taki układ uzyskanych rezultatów nie wynika z rzeczywistej rozbieżności między średnimi w dwóch populacjach, a ze zwykłej zmienności wyników (błędu próbkowania *sampling error*). Tak więc za istotnością statystyczną wyniku stoi przeszacowanie wielkości efektu. Nawet jeśli badacz ma rację i istnieje coś w danych, to rzadziej spotykane wyniki mają szansę przeskoczyć próg 0,05. Badacza mającego wrażenie, że odkrył coś właśnie na podstawie osiągniętych efektów, dotyka to, co w żargonie statystyków nazywa się *klątwą zwycięzcy* (Gelman, 2019)[3]. Wynik jest istotny statystycznie *ergo* możliwy do opublikowania, lecz ów sukces publikacyjny jest okupiony przeszacowaniem wielkości efektu. Klątwę zwycięzcy trudno zatem odczarować. Badacze, którzy opublikowali początkowo obiecujące rezultaty analiz, walczą o powtórzenie wyniku, a to nie następuje. Jednym z interesujących wyjaśnień jest powołanie się na efekt wygaszania, czyli zmniejszania się efektów wraz z upływem czasu. Klątwa zwycięzcy dotyka nie tylko samego badacza i jego zespół, ale i wszystkich, którzy opierają wiedzę na badaniach o przeszacowanych wynikach. Dzieje się tak dlatego, że patrzą oni na zjawisko przez pryzmat zniekształconych rezultatów.

Istotność statystyczna i (zbyt) wysoka moc testu

Istotny statystycznie wynik uzyskany w teście o dużej mocy wydaje się wymarzonym wynikiem analiz. Dlaczego? Byłby dowodem odkrycia zależności, a sam test

[3] Oto przykład badań, które najprawdopodobniej są efektem klątwy zwycięzcy – badania nad postawami mocy (*power pose;* Carney i in., 2010). W dużym skrócie: codzienne stawanie w pozycji widzianej w komiksach z superbohaterami ma zwiększać pewność siebie (więcej: Ranehill i in., 2015). Nauka już opuściła korytarze akademii i coraz więcej ludzi niezwiązanych z nią korzysta z analiz, promując wyniki. Na przykład jedna z popularniejszych trenerek fitness promuje codzienny trening z użyciem power pose, aby wzmocnić samoocenę (Lewandowska, 2018).

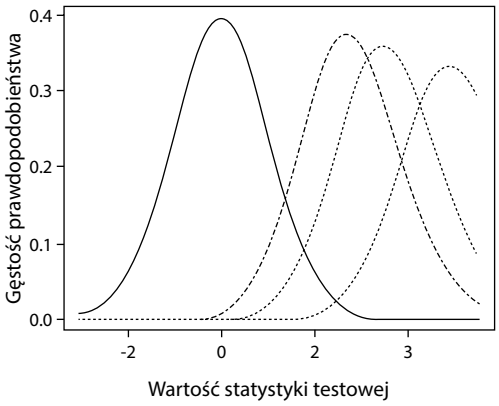
statystyczny – okazałby się zdolny do wykrycia efektu. Badacz mógłby wykorzystać dużą moc testu do przyjęcia, że hipoteza alternatywna została potwierdzona.

Spróbujmy wykazać błąd w stwierdzeniu hipotetycznego badacza, że istotny statystycznie wynik potwierdza hipotezę alternatywną, ponieważ test miał dużą moc. Pierwszy kontrargument, jaki nasuwa się, jest następujący: z czysto naukowego punktu widzenia żadne pojedyncze badanie nie jest w stanie udowodnić hipotezy badacza. Należałoby przeprowadzić replikacje i sprawdzić, czy uzyska się podobny wynik. Ale to argument, który nie zagłębia się w istotę analizy mocy. Wróćmy do momentu przed badaniami, a więc tam, gdzie, na etapie analizy mocy badacza postawiono przed zadaniem wymyślenia spodziewanej wielkości efektu, a wszystko po to, aby obliczyć potrzebną wielkość próby. Wówczas rozmytą hipotezę alternatywną (np. H_1: $\mu \neq 0$) zastąpiono przez konkretną postać (np. H_1: $\mu = 0{,}1$). Warto zadać pytanie, co w takim razie zostało potwierdzone przez istotny statystycznie wynik przy dużej mocy testu? Brak równości zawarty w oryginalnej hipotezie alternatywnej czy dokładna wartość wielkości efektu wstawiona podczas analizy mocy do wyliczenia liczby osób badanych? To pytanie można sformułować inaczej – czy badacz może powiedzieć, że istotny statystycznie wynik przy dużej mocy testu potwierdza, że wielkość efektu w populacji wynosi właśnie tyle, ile zaplanował podczas analiz mocy? Czy może jedynie dowieść, że istnieje jakikolwiek efekt (tu: $\mu = 0{,}1$), choćby infinitezymalny? To ważne pytanie, na które nie ma prostej odpowiedzi.

Rozważmy inny scenariusz, w którym istotny statystycznie wynik został uzyskany w teście o mocy znacznie przekraczającej próg 80%, np. 95% lub 99%. Odpowiada on temu, co już zostało nazwane badaniami o zbyt dużej mocy. Jak interpretować istotność statystyczną uzyskaną w testach o zbyt dużej mocy? Poznanie odpowiedzi na to pytanie wymaga przyjrzenia się związkowi między p-wartością a wielkością próby. Jest on dwojaki: nieistniejący przy prawdziwości hipotezy zerowej i ścisły przy prawdziwości hipotezy alternatywnej. Jeśli w populacji zależności między zmiennymi nie ma, to p-wartość i wielkość efektu są niezależne, wzrost wielkości próby nie wpływa na p-wartość. Wynik testu nie będzie jeszcze bardziej nieistotny statystycznie po zbadaniu większej liczby osób. Za to, gdy zależność w populacji istnieje, badacze obserwują, że istotność statystyczna wszystkich, nawet trywialnych efektów, rośnie wraz z rosnącą wielkością próby. Fakt, że duże próby często idą w parze z dużą mocą testu, jest źródłem obaw przed testami o zbyt dużej mocy (Utts, 2005; Brzeziński, 1997, s. 337). Wkrótce się okaże, że zjawisko to nie jest paradoksem statystycznym, lecz problemem w interpretacji istotności statystycznej w teście o bardzo dużej mocy.

Gdy patrzymy na rysunek 5, staje się jasne, dlaczego za wzrostem próby idzie istotność statystyczna. Załóżmy, że badacz spodziewa się małej wielkości efektu. Rysunek 5 (s. 95) jest podobny do rysunku 3, lecz różni się wielkością próby. Liczba obserwacji rośnie, a położenie i kształt rozkładu statystyki testowej znów ulega zmianie i krzywe oddalają się od siebie. Przy czym zwiększenie wielkości próby nie zmieniło wielkości efektu w populacji – badacz nadal oczekuje *d* Cohena równej 0,5.

Wraz ze zmieniającą się liczebnością próby, rozkład statystyki testowej dla hipotezy alternatywnej również zmienia się właśnie przez odsuwanie się od rozkładu statystyki testowej dla hipotezy zerowej. To zjawisko rozsuwania się



Rysunek 5. Stopniowe przesuwanie się w prawo niecentralnego rozkładu statystyki testowej mimo stałości wielkości efektu d Cohena równej 0,5

Nota. Rosnąca wielkość próby działa na rozkład statystyki testowej zarówno przy prawdziwości hipotezy zerowej, jak i alternatywnej. Rozkład rozkład statystyki testowej dla prawdziwej hipotezy zerowej wysmukla się, lecz pozostaje w miejscu. Rozkład statystyki testowej dla hipotezy alternatywnej zmienia się zarówno pod względem kształtu, jak i położenia – wyłącznie dlatego, że zmienia się wielkość próby, która ma wpływ na parametr niecentralności. Właśnie to zjawisko jest odpowiedzialne za złudzenie istotności statystycznej wszystkich efektów (nawet tych trywialnych). Źródło: opracowanie własne.

krzywych będziemy obserwować ilekroć w populacji zaistnieje jakikolwiek efekt, co jest równoznaczne z tym, że prawdziwa wartość parametru jest niezerowa.

Wydaje się, że konstrukcja klasycznych testów statystycznych skazuje badaczy na to zjawisko, którego się obawiają, tj. rosnącej istotności statystycznej testu przy rosnącej wielkości próby. Tymczasem rosnąca wielkość próby zmienia rozkład statystyki testowej, tak jak zmienia ją rosnąca wielkość efektu.

Jednak obawa przed dużymi próbami, czy zbyt dużą mocą testu, ma racjonalne podstawy tylko wówczas, gdy istotność statystyczną zrówna się z istotnością rzeczywistą. Jeśli badacze będą się skłaniać do właśnie takiej interpretacji (m.in. Goodman, 2008), to również będą obawiać się dużych prób. Z pomocą ocenie ważności zjawiska w badaniach o dużych próbach przychodzi wielkość efektu, której zaletą jest niezależność od liczby zbadanych osób. W przeciwnym wypadku uznać należałoby, że przebadanie dużej liczby osób jest złym rozwiązaniem.

Nieistotność statystyczna i niska moc testu

W poprzednim paragrafie zobaczyliśmy, że niska moc testu stanowi utrudnienie w interpretacji istotnego statystycznie wyniku. Jednak badacz, który otrzyma wynik nieistotny statystycznie, nie będzie w lepszej sytuacji. Jeśli test ma małą moc, np. 20%, wówczas 80% jego wyników będzie znajdowało się w regionie odpowiadającym akceptacji hipotezy zerowej.

Przyglądając się rysunkowi 3, zauważyć można, że przy niskiej mocy testu krzywe rozkładów statystyki testowej pod warunkiem prawdziwości hipotezy zerowej i alternatywnej leżą blisko siebie. Czasem zbyt blisko, by stwierdzić, czy wynik jest nieistotny statystycznie, ponieważ hipoteza zerowa jest prawdziwa, a test ma zbyt małą moc do wykrycia efektu. Wprowadzenie obowiązku zapewnienia odpowiedniej wysokości próby wytrąciłoby stosowane wówczas usprawiedliwienie: „gdyby było więcej osób, to pojawiłaby się istotność statystyczna".

Kłopot z takimi wynikami jest dwojaki. Po pierwsze, najczęściej odkładane są do szuflady, pogłębiając problem o tej samej nazwie (*file drawer problem*). Kryzys replikacyjny to nie tylko efekt testów niskiej mocy, to również problem ze zniekształconym obrazem badanego zjawiska. Literatura dysponuje przede wszystkim istotnymi statystycznie wynikami, a te, które nie przeskoczyły progu 5-procentowego, nie są publikowane, odkładane do szuflady. Ten aspekt dotyczy nieistotnych statystycznie wyników zarówno w testach niskiej mocy, jak i wysokiej mocy.

Drugi kłopot jest taki, że niektóre badania są z góry skazane na niską moc i nieistotność statystyczną. To badania z obszarów, w których trudno o obserwacje (badania z udziałem zwierząt, przypadki rzadkich chorób), a spodziewane wielkości efektu są małe. Badacz albo otrzyma mało informacyjny wynik nieistotny statystycznie, albo ryzykuje klątwą zwycięzcy. Badacz, który nie może sobie pozwolić na przebadanie dużej liczby osób, bo fizycznie one nie istnieją na świecie, może oprzeć się na wielkościach efektu lub na nieklasycznej statystyce, np. bayesowskiej.

Nieistotność statystyczna i wysoka moc testu

Czwarty scenariusz to połączenie nieistotności statystycznej wyniku (np. p = ,32) i wysokiej mocy testu (np. 90%). *Skoro mój test miał dużą moc* – rozumuje badacz – *to był zdolny do wykrycia zamierzonego efektu. Mój wynik jest nieistotny statystycznie, więc hipoteza zerowa o braku efektu musi być prawdziwa*. Wysoka moc testu uwiarygodniałaby wnioski płynące z nieistotności statystycznej.

Wysoką moc statystyczną testu statystycznego łatwo pomylić z czułością testu medycznego. Mimo wszystko, test statystyczny rządzi się innymi prawami niż test medyczny. Aby zobaczyć, co mogłoby stać za nieistotnością statystyczną wyniku, musimy cofnąć się do chwili, w której badacz wybiera spodziewaną wielkość efektu w celu wykonania analizy mocy. Nigdy nie ma pewności, że prawidłowo odgadł wielkość efektu w populacji. Może ją przeszacować i wybrać większą wartość niż takowa istnieje. Przykładowo, populacyjna wielkość efektu wynosi 0,57 sekundy. Tego badacz nie wie, bo jest to ta wartość, którą ma odgadnąć. Spodziewa się zatem wyższej wartości równej jednej sekundzie, zawyżając wielkość efektu. W konsekwencji, wykorzystując analizę mocy, wyznaczy mniejszą liczbę osób, dostosowaną do wyższej wartości, zamiast postąpić na odwrót. Innymi słowami, niechcąco *dostosuje ostrość do belki, szukając drzazgi*. To z kolei zwiększy błąd II-go rodzaju, a test przestanie być mocny na zakładanym poziomie 80% i nieświadomy badacz znajdzie się w polu scenariusza trzeciego.

PODSUMOWANIE

Kryzys replikacyjny pojawił się w szeroko pojętej nauce i dotyka każdej dziedziny, której dotyczy. Naukowcy znaleźli na niego sposób i zaczęli stosować analizę mocy. Zainteresowanie nią wzrosło również dlatego, że coraz częściej redakcje czasopism naukowych wymagają uzasadnienia takiej, a nie innej wielkości próby. Prawidłowo przeprowadzona analiza mocy odbywa się przed wykonaniem badania (Nakagawa, Foster, 2004). Trudno jednak poprzestać na wyliczeniu wielkości próby bez odnoszenia się do istotności statystycznej po uzyskaniu wyników. Tymczasem istotność statystyczna i analiza mocy to dwie odrębne kwestie. Istotność statystyczna jest miarą dopasowania danych do hipotezy zerowej (Wasserstein, Lazar, 2016). Natomiast analiza mocy służy do kontroli długofalowego popełniania błędów II-go rodzaju (Neyman, 1977). Stosując ją, pozostajemy w obszarze interpretacji frekwentystycznej – badacz nie wie, czy jego wynik jest jednym z odsetka błędnych wyników, ponieważ jedynie kontroluje – w dłuższej perspektywie – omyłkowe odrzucenie prawdziwej hipotezy alternatywnej.

Interpretacja wyników istotnych bądź nieistotnych statystycznie na tle małej lub dużej wielkości efektu prowadzi do problemów. W tym artykule rozważyliśmy cztery scenariusze, a w żadnym z nich nie ma jednoznacznej instrukcji interpretacji. Wnioskowanie o prawdziwości hipotezy zerowej na podstawie p-wartości i poziomu błędu II-ego rodzaju zdaje się być kolejną – po interpretacji istotności statystycznej w kategoriach błędu I-go rodzaju (Hubbard i Bayarri, 2003) – próbą integracji niekompatybilnych szkół myśli statystycznej.

BIBLIOGRAFIA

Brzeziński, J. (1997). *Metodologia badań psychologicznych*. Wydawnictwo Naukowe PWN.

Carney, D. R., Cuddy, A. J. C., Yap, A. J. (2010). Power Posing: Brief Nonverbal Displays Affect Neuroendocrine Levels and Risk Tolerance. *Psychological Science*, *21*(10), 1363–1368. DOI: https://doi.org/10.1177/0956797610383437.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed). L. Erlbaum Associates.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, *45*(12), 1304–1312. https://doi.org/10.1037/0003-066X.45.12.1304.

Cumming, G. (2011). Understanding The New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis. Routledge. ISBN 9780415879682.

Gelman, A. (2019, January 4). Yes, it makes sense to do design analysis ("power calculations") after the data have been collected. Statistical Modeling, Causal Inference, and Social Science.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, 33(5), 587–606. DOI: https://doi.org/10.1016/j.socec.2004.09.033.

Hubbard, R., Bayarri, M. J. (2003). Confusion Over Measures of Evidence (p 's) Versus Errors (α 's) in Classical Statistical Testing. *The American Statistician*, *57*(3), 171–178. DOI: https://doi.org/10.1198/0003130031856.

Huberty, C. J. (1993). Historical Origins of Statistical Testing Practices: The Treatment of Fisher versus Neyman-Pearson Views in Textbooks. *The Journal of Experimental Education*, *61*(4), 317–333. DOI: http://www.jstor.org/stable/20152384.

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLoS Medicine*, *2*(8), e124. DOI: https://doi.org/10.1371/journal.pmed.0020124.

Jarmakowska-Kostrzanowska, L. (2016). W statystycznym matriksie: kontrowersje wokół testowania istotności hipotezy zerowej (null hypothesis significance testing, NHST) oraz p-wartości. *Psychologia Społeczna*, *4*(39), 458–473. DOI: https://doi.org/10.7366/1896180020163906.

Kelley, K. (2013). Effect Size and Sample Size Planning. W: *The Oxford Handbook of Quantitative Methods* (Vol. 1, pp. 206–222).

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, Š., Bernstein, M. J., … Nosek, B. A. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology, 45*(3), 142–152. DOI: http://dx.doi.org/10.1027/1864-9335/a000178.

Lewandowska, A. (2018, February 18). Power pozycja! Healthy Plan by Ann. https://hpba.pl/power-pozycja/.

Mayo, D. G. (2018). *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars* (1st ed.). Cambridge University Press. DOI: https://doi.org/10.1017/9781107286184.

Meyer, A., Frederick, S., Burnham, T. C., Guevara Pinto, J. D., Boyer, T. W., Ball, L. J., Pennycook, G., Ackerman, R., Thompson, V. A., Schuldt, J. P. (2015). Disfluent fonts don't help people solve math problems. *Journal of Experimental Psychology: General*, *144*(2), e16– e30. DOI: https://doi.org/10.1037/xge0000049.

Murphy, K.R., Myors, B., i Wolach, A. (2014). *Statistical Power Analysis: A Simple and General Model for Traditional and Modern Hypothesis Tests*. Routledge.

Nakagawa, S., Foster, T. M. (2004). The case against retrospective statistical power analyses with an introduction to power analysis. *Acta Ethologica*, *7*(2), 103–108. DOI: https://doi.org/10.1007/s10211-004-0095-z.

Neyman, J. (1977). Frequentist Probability and Frequentist Statistics. *Synthese*, *36*(1), 97–131. JSTOR. DOI: 10.1007/BF00485695.

Ranehill, E., Dreber, A., Johannesson, M., Leiberg, S., Sul, S., Weber, R. A. (2015). Assessing the Robustness of Power Posing: No Effect on Hormones and Risk Tolerance in a Large Sample of Men and Women. *Psychological Science*, 26(5), 653–656. DOI: https://doi.org/10.1177/0956797614553946.

Sirota, M., Theodoropoulou, A., Juanchich, M. (2020). Disfluent fonts do not help people to solve math and non-math problems regardless of their numeracy. *Thinking & Reasoning*, 1– 18. https://doi.org/10.1080/13546783.2020.1759689.

Strack, F., Martin, L. L., & Stepper, S. (1988). Inhibiting and facilitating conditions of the human smile: A nonobtrusive test of the facial feedback hypothesis. *Journal of Personality and Social Psychology, 54*(5), 768–777. DOI: https://doi.org/10.1037/0022-3514.54.5.768.

Utts, J. M. (2005). *Seeing through statistics* (3rd ed). Thomson, Brooks/Cole.

Wagenmakers, E.-J., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams, R. B., Albohn, D. N., Allard, E. S., Benning, S. D., Blouin-Hudon, E.-M., Bulnes, L. C., Caldwell, T. L., Calin-Jageman, R. J., Capaldi, C. A., Carfagno, N. S., Chasten, K. T., Cleeremans, A., Connell, L., DeCicco, J. M., … Zwaan, R. A. (2016). Registered Replication Report: Strack, Martin, Stepper (1988). *Perspectives on Psychological Science*, *11*(6), 917–928. DOI: https://doi.org/10.1177/1745691616674458.

Wasserstein, R. L., Lazar, N. A. (2016). The ASA Statement on p-Values: Context, Process, and Purpose. The American Statistician, *70*(2), 129–133. DOI: 10.1080/00031305.2016.1154108.

Wolski, P. (2017). Istotność statystyczna III. Od rytuału do myślenia statystycznego. *Rocznik Kognitywistyczny*, *9*(2016). DOI: https://doi.org/10.4467/20843895RK.16.007.6413.