Rocznik Kognitywistyczny 9/2016, s. 71–85
doi: 10.4467/20843895RK.16.007.6413
www.ejournals.eu/Rocznik-Kognitywistyczny

PIOTR WOLSKI

Uniwersytet Jagielloński
Instytut Psychologii

Istotność statystyczna III. Od rytuału do myślenia statystycznego

Statistical significance III. From ritual to statistical thinking

Abstract: One of the more prominent problems of significance testing is ritualisation of their practical use and interpretation. In the present, third part of the series, reasons and manifestations of that rigidity have been discussed, and an alternative, sometimes labeled "statistical thinking", presented. Matching a statistical significance testing scenario to the needs of the specific research program constitutes a part of statistical thinking. Some typical scenarios have been described, with the intent of showing how the same statistical tool, depending on it's assumptions, can have differing use in research.

Keywords: statistical inference, null hypothesis significance testing, NHST, *p*-value, statistical power

W poprzednich dwóch częściach cyklu poświęconego istotności statystycznej opisane zostały typowe problemy, jakie napotykają psychologowie próbujący zrozumieć logikę wnioskowania statystycznego oraz, odpowiedzialne za część trudności interpretacyjnych, ograniczenia testów istotności. Narastająca świadomość tych niedostatków [np. Cohen, 1994; Gigerenzer, 1998] skłoniła radę naukową Amerykańskiego Towarzystwa Psychologicznego (APA) do powołania pod koniec ubiegłego wieku komisji znanych ekspertów, specjalistów z zakresu metodologii badań psychologicznych, z zadaniem ustosunkowania się do podnoszonych przez krytyków kwestii i opracowania zestawu wytycznych dla badaczy publikujących w periodykach Towarzystwa. Efekty tych prac ogłoszono w piśmie „American Psychologist" [Wilkinson, APA Task Force on Statistical Inference, 1999]. Komisja postuluje zastąpienie mechanicznych zero-jedynkowych decyzji w sprawie testowanych hipotez bardziej subtelnym różnicowaniem wyników. Informację o spełnieniu (lub nie) kryterium $p \leq 0{,}05$ winno zastąpić – zgodnie z zaleceniem Fishera – podanie obliczonej wartości prawdopodobieństwa p, np. $p = 0{,}012$. Eksperci proponują szersze stosowanie estymacji przedziałowej i badanie wielkości efektów, a także wskazują na potrzebę stosowania najprostszych adekwatnych do problemu analiz, w tym oszczędne stosowanie obniżających moc statystyczną porównań wielokrotnych. Poza kwestiami związanymi z wnioskowaniem statystycznym zwracają też uwagę na wiele innych spraw ważnych w projektowaniu, analizie i opisie badań. Odnoszą się do rzeczy fundamentalnych, takich jak warunki zasadności

wnioskowania przyczynowego, ale i drobnych, na przykład zalecają, by przytaczając wyniki, ograniczać się do tylu miejsc po przecinku, ile uzasadnia faktyczna dokładność pomiaru, a nie kopiować mechanicznie wydruków generowanych przez program statystyczny.

Opracowany w ramach bezprecedensowej inicjatywy APA raport to zestaw cennych wskazówek, ale też ważka deklaracja potrzeby częściowej rewizji dotychczasowej praktyki metodologicznej. Przybywa głosów nawołujących do takich zmian. Poczytne pismo amerykańskiego Towarzystwa Nauk Psychologicznych, „Psychological Science", rekomenduje swoim autorom, podobną do zaleceń ekspertów APA, ale sformułowaną bardziej kategorycznie, propozycję Geoffa Cumminga ochrzczoną przezeń – chyba nieco na wyrost – mianem „nowej statystyki" [Cumming, 2014]. Cumming zwraca uwagę na wady wrodzone testów istotności, zwłaszcza na zwodniczość samego pojęcia istotności statystycznej, którego najlepiej byłoby jego zdaniem unikać. Postuluje zastępowanie, tam gdzie to możliwe, testów istotności przedziałami ufności, przy czym proponuje zaczynać już na etapie stawiania pytań badawczych, które lepiej formułować w kategoriach estymacji i wielkości efektów niż sztucznie zdychotomizowanych hipotez. Metodolog zachęca do myślenia w kategoriach metaanalitycznych i – podobnie jak Fisher – podkreśla znaczenie replikacji.

Zdecydowanie najdalej poszli redaktorzy pisma „Basic and Applied Social Psychology" (BASP), którzy uznawszy obie tradycyjne metody wnioskowania statystycznego, tj. testy istotności oraz przedziały ufności, za błędne, ogłosili zakaz ich stosowania w pracach zgłaszanych do druku w tym periodyku [Trafimow i Marks, 2015]. Zamiast tego wzywają autorów do stosowania „mocnej statystyki opisowej". W połowie lat dziewięćdziesiątych ubiegłego wieku redaktor naczelny „Memory and Cognition", Geoffrey Loftus, podjął mniej radykalną, ale podobną próbę zmiany standardów, zachęcając autorów do szerszego korzystania z opisu statystycznego oraz zastąpienia weryfikacji hipotez estymacją przedziałową. Środowisko nie było jednak na tę zmianę gotowe [por. Finch i in., 2004] i po ustąpieniu Loftusa nowa redakcja przywróciła wcześniejsze *status quo*. Czy obecna, idąca znacznie dalej, inicjatywa redaktorów BASP ma większe szanse powodzenia? Z jednej strony pewnie tak, bowiem świadomość wad tradycyjnego wnioskowania statystycznego, a tym samym i potrzeba zmian, jest dziś bardziej powszechna. Do tego bezkompromisowość stanowiska redakcji prowokuje żywy oddźwięk, poszerzając dyskusję nad standardami statystyki stosowanej w badaniach psychologicznych[1]. Z drugiej strony całkowity zakaz wnioskowania statystycznego nosi znamiona wylewania dziecka z kąpielą. Poznanie nomotetyczne – odkrywanie praw ogólnych – przyjmuje w naukach indukcyjnych postać wnioskowania o populacji na podstawie obserwacji poczynionych w próbie. Zakaz wnioskowania statystycznego oczywiście nie likwiduje potrzeby szacowania para-

[1] Decyzję BASP skomentował przewodniczący Amerykańskiego Towarzystwa Statystycznego, który zapowiedział opracowanie przez grupę ekspertów Towarzystwa raportu poświęconego problemom związanym ze stosowaniem i interpretacją metod inferencyjnych. Zaapelował przy tym do redaktorów pisma oraz wszystkich, którzy podzielają ich wątpliwości, „by nie odrzucali właściwie i adekwatnie stosowanych metod wnioskowania statystycznego" [Wasserstein, 2015]. Przeciwko zakazowi testów istotności wypowiedział się także przewodniczący oraz kilku prominentnych członków brytyjskiego Królewskiego Towarzystwa Statystycznego [StatsLife, 2015]. Komentarz opublikowało prestiżowe „Nature" [Woolston, 2015].

metrów populacji, ale paradoksalnie pozbawia badacza służących temu szacowaniu narzędzi. Owszem, są one, jak starałem się pokazać w poprzednich częściach cyklu, obciążone, niewłaściwie rozumiane, a w konsekwencji także nie najlepiej stosowane, ale bez nich badaczowi zostają tylko, bardzo zawodne [Kahneman, 2011; Tversky i Kahneman, 1971], intuicje statystyczne. Upowszechnianie wiedzy o ograniczeniach tradycyjnych metod inferencyjnych i sposobach maksymalizowania ich rzetelności wydaje się rozwiązaniem bardziej adekwatnym niż opcja zerowa. Tym bardziej, że ta ostatnia nie rozwiązuje najważniejszej kwestii, która w praktyce ciąży na jakości analizy statystycznej bardziej niż ograniczenia teoretyczne: rytualizacji [Gigerenzer, 2004].

Schemat

Wiele lat temu, jako jeden z beneficjentów programu stypendialnego oferowanego przez kolegia oksfordzkie badaczom zza – korodującej już – żelaznej kurtyny, postanowiłem wykorzystać nadarzającą się okazję do nadrobienia luk w wykształceniu statystycznym. Od bardziej doświadczonych angielskich kolegów dowiedziałem się, że pozycją kanoniczną w zakresie analizy statystycznej jest podręcznik Rogera E. Kirka, ale ponieważ to dzieło opasłe i dla przeciętnego psychologa mało zrozumiałe, na co dzień wygodniej korzystać z popularnego podręcznika statystyki dla uniwersytetu otwartego [Greene i D'Oliveira, 1982]. Niegruby zeszyt formatu A4 wyglądał zachęcająco: wydawca informował, że autorka uzyskała doktorat z psychologii na Uniwersytecie Oksfordzkim, cytował też entuzjastyczną rekomendację biuletynu Brytyjskiego Towarzystwa Psychologicznego. Wstęp do pierwszej części, zatytułowanej *Statystyka nieparametryczna*, uspokajał, że z opisanych w niej metod można bezpiecznie korzystać, nie rozumiejąc, co znaczy tytuł, ale jeśli komuś ta kwestia nie daje spokoju, znajdzie stosowne wyjaśnienie we wstępie do części drugiej, pod tytułem *Statystyka parametryczna*. Na wewnętrznej stronie okładki przedstawiono schemat doboru testów do problemów badawczych. Użytkownik odpowiadał na kolejne pytania: Czy interesują go korelacje, czy różnice między średnimi? Czy w badaniu występuje jedna, czy więcej zmiennych? Ile jest warunków eksperymentalnych? Czy badano te same lub dobierane parami osoby, czy różne? Zależnie od odpowiedzi przejrzysty diagram prowadził do właściwego testu, a ściślej – do pary testów: jednego nieparametrycznego i jednego parametrycznego. Zasadniczą część poradnika stanowiły opisy testów – z prostymi wzorami pozwalającymi łatwo obliczyć wartości statystyk testowych. Podpowiadały też, jak wyznaczyć liczby stopni swobody, co umożliwiało odczytanie z zamieszczonych na końcu tablic statystycznych uzyskanego poziomu istotności. Całość poprzedzał wstęp, nadzwyczaj krótko i jasno omawiający najważniejsze pojęcia teoretyczne. Tam, gdzie prostota i ścisłość wchodziły w konflikt, wygrywała raczej ta pierwsza, więc można było przeczytać na przykład, iż testy dostarczają „prawdopodobieństwa, że (…) wyniki są istotne" (*sic!*), ale czynią to w sposób, „który może się wydać nieco paradoksalny" [Greene i D'Oliveira, 1982, s. 31]. Nie wgłębiając się w zawiłości teoretyczne, z radością skorzystałem z rekomendacji, kupiłem na wyprzedaży dobry kalkulator, przećwiczyłem przykłady, przerysowałem diagram decyzyjny i wróciłem

do kraju z miłym – ale niestety słabo uzasadnionym – wrażeniem, że oto nauczyłem się w Oksfordzie statystyki.

Popularność opisanego podręcznika nie była przypadkowa (dziś można kupić jego kolejne, trzecie już wydanie). Doskonale odpowiadał na zapotrzebowanie, prezentując w klarownej, skondensowanej formie prosty przepis pomagający autorom prac empirycznych dopasować odpowiednie testy istotności do typowych sytuacji badawczych. Opanowanie opisanego tam algorytmu wydawało się istotą praktycznej kompetencji statystycznej psychologa doświadczalnika. W oczach środowiska za eksperta uchodził ten, kto posiadał szczegółową wiedzę na temat detali popularnych procedur, na przykład rozróżniał wiele różnych rodzajów testów *post hoc*, wiedział, że ANOVA z powtarzanymi pomiarami wymaga spełnienia założenia sferyczności, umiał w razie potrzeby zastosować poprawkę Greenhouse'a-Geissera, albo wiedział, że poprawka Bonferroniego czyni test nadmiernie konserwatywnym. Początkujący badacze podpytywali starszych kolegów, jakiego testu powinni użyć, ale raczej nikomu nie przychodziło do głowy pytać, czy w ogóle używać testu istotności, a jeśli tak, to po co, albo jak interpretować wyliczoną wartość *p*. Odpowiedzi wydawały się oczywiste: test potrzebny jest zawsze, bo stanowi statystyczne potwierdzenie realności wyniku. W kwestii interpretacji lektura doniesień z badań też nie zostawiała wiele wątpliwości – efekty, dla których $p \leq 0{,}05$, od tych, dla których *p* było pechowo większe od magicznego kryterium, zdawała się oddzielać głęboka przepaść. Mój kolega – wtedy zdolny doktor, dziś profesor matematyki – spytał mnie w tamtym okresie, jakich poziomów istotności używa się w psychologii. Zdziwił się, że minimum stanowi 0,05, bo uważał, że zważywszy na przedmiot, właściwsze byłoby stosowanie poziomu 0,1 albo nawet 0,2. Brzmiało to jak podwójne bluźnierstwo – nie dość, że kryterium jest tak niskie, to jeszcze badacz miałby je dobierać sam!

Mimo nieporównanie lepszych warunków pracy naukowej, dostępności literatury fachowej i specjalistycznych narzędzi, dziś wciąż naczelnym problemem zastosowań wnioskowania statystycznego w psychologii nie są ograniczenia testów istotności, ale podobny do opisanego, mechaniczny, sztywny sposób ich stosowania i interpretacji. Badacz wypełniający „rytuał istotności" (*null ritual* [Gigerenzer, 2004]) sprawdza, jaki test stosuje się w danej sytuacji, przeprowadza obliczenia, po czym uznaje testowany efekt za realny, jeśli $p \leq 0{,}05$. W przeciwnym wypadku stwierdza, mniej lub bardziej wprost, że efekt w populacji nie istnieje. Z reguły nie zna mocy testu, a wielkość próby dobiera na oko, wzorując się na innych podobnych badaniach. Przyjmuje też zwykle, że efekt istotny statystycznie jest zarazem znaczący praktycznie i nawet nie próbuje szacować jego wielkości.

Myślenie statystyczne

Rytuał istotności obsadza test w roli autonomicznego arbitra realności i znaczenia obserwowanych efektów empirycznych. Ten powstał jednak jako narzędzie jedynie pomocnicze. Ocena rzetelności efektów nie może być mechaniczna, bowiem żadna jakościowa różnica nie dzieli tych, dla których ryzyko błędu pierwszego rodzaju wynosi 0,05, od tych, przy których jest ono równe 0,06. Także wybór kryterium

$\alpha = 0{,}05$ jest arbitralny. Nie ma podstaw, by twierdzić, że nauka rozwija się optymalnie akurat wtedy, gdy fałszywe odrzucenia hipotez zerowych zdarzają się raz na 20, a nie 15 czy 112 przypadków. Podobnie nieuzasadniona jest, typowa dla zrytualizowanego wnioskowania statystycznego, automatyczna akceptacja efektów spełniających warunek $p \leq 0{,}05$ niezależnie od celu i kontekstu analizy: zarówno w skromnych badaniach pilotażowych, jak i w badaniach na olbrzymiej próbie; tak samo przy wysokim, jak przy niskim apriorycznym prawdopodobieństwie istnienia testowanego efektu; wtedy gdy błędna decyzja może mieć poważne konsekwencje praktyczne i wtedy gdy cierpi jedynie ego badacza. Choć łatwo jest sporządzić listę typowych grzechów zrytualizowanego testowania istotności (większość omówiliśmy wcześniej), podobne scharakteryzowanie właściwej alternatywy przychodzi trudniej – jej istotą jest bowiem brak jednego schematu.

Pożądanym celem szeroko rozumianej edukacji statystycznej wydaje się „myślenie statystyczne" [Abelson, 1995; Gigerenzer, 1998; Ostasiewicz, 2012; Palij, 2012], które można z grubsza zdefiniować jako *umiejętność adekwatnego użycia narzędzi statystycznych przy tworzeniu narracji naukowej odwołującej się do danych empirycznych*[2]. Właściwie prowadzona analiza statystyczna, choć korzysta ze zalgorytmizowanych procedur, sama ma charakter otwarty dzieła autorskiego, wymagającego decyzji, których nie da się scedować na algorytm. Brzeziński [2012b] słusznie podkreśla, że analiza statystyczna jest wtórna w stosunku do teorii i planu badawczego. Związek tych elementów jest nawet jeszcze ściślejszy, bowiem plan badawczy opracowuje się zwykle, mając na względzie możliwe strategie przyszłej analizy, a teoria nie tylko wyznacza działania na poziomie operacyjnym, ale też sama jest przez nie zwrotnie kształtowana. Fisher pisał: „Procedura statystyczna i plan eksperymentalny to tylko dwa różne aspekty tej samej całości"[3]. Będąc częścią działania, nazywanego nie bez powodu *twórczością* naukową, analiza danych wymaga sporej dozy myślenia dywergencyjnego. Nie ogranicza się jednak do samych problemów otwartych – w odniesieniu do testowania istotności kluczowy jest wybór i konsekwentna realizacja jednego z kilku dostępnych scenariuszy.

Scenariusze testowania istotności

W toku analizy statystycznej nie tylko dokonuje się wyboru testu statystycznego, ale też przyjmuje rozstrzygnięcia bardziej podstawowej natury, dotyczące celu i sposobu użycia owego testu. Można w tym zakresie wyróżnić kilka mniej lub bardziej typowych schematów, które nazywam dalej scenariuszami.

[2] Brzeziński [2012a; 2012b] używa w podobnym, ale nieco węższym znaczeniu pokrewnego pojęcia „świadomość metodologiczna".

[3] „Statistical procedure and experimental design are only two different aspects of the same whole (…)" [Fisher, 1971, s. 3].

Wstępna selekcja

W scenariuszu selekcyjnym zadaniem testu istotności jest odrzucenie tych efektów, których rzetelność jest zbyt mała, by nawet w słabym trybie przypuszczającym dało się o nich powiedzieć coś znaczącego. Taki cel testu sugeruje Fisher, w cytowanym w poprzedniej części fragmencie pisząc o gotowości ignorowania wyników niespełniających przyjętego standardu istotności i „eliminowania w ten sposób większości fluktuacji wywołanych czynnikami losowymi" (Fisher, 1971, s. 13–14). Scenariusz selekcyjny zakłada automatyzm działania – mechaniczne oczyszczenie pola zainteresowania badacza z efektów, które nie są warte jego uwagi. Aby nie ryzykować wylania dziecka z kąpielą, tj. odrzucenia efektów, które choć obecne w populacji, ze względu na ograniczenia mocy statystycznej nie przekroczyłyby bardziej rygorystycznego kryterium, należy stosować raczej liberalne poziomy istotności, nawet rzędu $\alpha = 0{,}1$ lub 0,2.

Użycie selekcyjne testu istotności można luźno porównać do zaaplikowania filtru szumu w procesie obróbki fotografii cyfrowej. Inaczej niż przy ręcznym retuszu, kiedy fotograf sam wyszukuje i usuwa indywidualne zakłócenia w obrazie, filtr szumu automatycznie identyfikuje i eliminuje te elementy obrazu, które spełniają ustalone przez fotografa kryterium progowe. Właściwy dobór owego kryterium decyduje o efektywności filtru: przy kryterium zbyt liberalnym filtr słabo usuwa szum, przy zbyt konserwatywnym – usuwa razem z nim ważne detale.

Scenariusz selekcyjny ma oczywiście największy sens na wstępnym etapie analizy. Trzeba tu szczególnie pamiętać o różnicy między istotnością statystyczną a praktyczną: jedyny przywilej, jaki w owym scenariuszu zyskują wyniki statystycznie istotne – czytaj: wystarczająco niezgodne z hipotezą zerową – to warunkowe uniknięcie kosza na śmieci. Sformułowanie mocniejszych wniosków na temat populacji wymaga dalszej analizy statystycznej, teoretycznej i – co najważniejsze – mocniejszych danych empirycznych, czyli przekonujących replikacji [Wojciszke, 2004] oraz metaanaliz.

Konfirmacja

Scenariusz konfirmacyjny[4] dotyczy sytuacji, w której badacz testuje przewidywania prawdopodobne w świetle rozważań teoretycznych, zwykle uwiarygodnionych dodatkowo przez inne wyniki empiryczne. Kategoria ta obejmuje także replikacje wcześniejszych pozytywnych wyników. Choć trudno to jednoznacznie ocenić, praktyka badawcza zdaje się sugerować, że w typowych badaniach konfirmacyjnych powszechnie stosowane kryterium istotności $\alpha = 0{,}05$ sprawdza się względnie dobrze. Prawdopodobieństwo błędu pierwszego rodzaju wynosi przy tym kryterium 1/20, co może się wydawać znaczną wartością, jednak – jak zostało szerzej wyjaśnione w poprzednich dwóch częściach niniejszego cyklu – wbrew pozorom nie odpowiada ryzyku, że statystycznie istotny wynik okaże się fałszywym alarmem. Zarówno wartość α, jak i p opisują prawdopodobieństwo zaobserwowania danych uzasadniających odrzuce-

[4] Użyty tu termin „konfirmacja" nie odnosi się do logiki weryfikacji hipotez statystycznych, bo ta ma charakter falsyfikacyjny. Chodzi o rozróżnienie pomiędzy dopasowywaniem danych empirycznych do modelu teoretycznego, tworzonego *a priori* (konfirmacja) lub *a posteriori* (eksploracja).

nie hipotezy zerowej, *gdy* ta jest prawdziwa, $P(D|H_0)$, a nie powiązane z nim, ale nie tożsame, prawdopodobieństwo, *że* owa hipoteza jest prawdziwa, $P(H_0|D)$[5]. To ostatnie można oszacować za pomocą formuły Bayesa, wyrażającej prawdopodobieństwo warunkowe $P(H_0|D)$ jako iloczyn prawdopodobieństwa błędu pierwszego rodzaju $P(D|H_0)$ oraz ilorazu prawdopodobieństw *a priori*, $P(H_0)/P(D)$. Z owej zależności wynika, że im mniej prawdopodobna *a priori* jest hipoteza zerowa, tym mniejsze prawdopodobieństwo, że zaobserwowany w próbie efekt, spełniający kryterium istotności $p \leq \alpha$, jest w istocie tylko przypadkowym artefaktem. Wynik o tym samym poziomie istotności zaobserwowanej p jest mocniejszym dowodem w badaniach konfirmacyjnych niż eksploracyjnych, ponieważ w tych pierwszych badacz ma podstawy do oczekiwania go *a priori*. Wartość dowodowa jest oczywiście tym większa, im mocniejsze są, uzasadniające owo oczekiwanie, przesłanki teoretyczne i empiryczne. Trzeba też zauważyć, że choć tradycyjne kryterium $\alpha = 0{,}05$ definiuje rozsądny poziom minimalnej wartości dowodowej, jest to jednakowoż poziom *minimalny* – wyższą wagę mają wyniki spełniające bardziej rygorystyczne kryteria, a prawdziwie dużą dopiero te, które udaje się systematycznie replikować.

Eksploracja

Podczas gdy w badaniach konfirmacyjnych obserwacje empiryczne są dodatkowo uwiarygodnione przez wcześniejsze przewidywania, w scenariuszu eksploracyjnym znajdują one tylko daleko słabsze wsparcie w postaci wyjaśnień *a posteriori*, a bywa że nie mają nawet takiej walidacji. Standardowe kryterium istotności $\alpha = 0{,}05$ wydaje się w takim przypadku zbyt liberalne. Ryzyko błędu pierwszego rodzaju równe 1/20 definiuje poziom wiarygodności dowodu, który nie jest radykalnie większy niż w przypadku, sugerowanego wcześniej jako właściwe dla scenariusza selekcyjnego, ryzyka 1/10, czy nawet 1/5[6]. We wszystkich tych przypadkach nieoczekiwany pozytywny wynik testu jest wystarczającym uzasadnieniem właściwie tylko dla sformułowania hipotezy roboczej, weryfikowanej w kolejnych badaniach.

Trudno sugerować, jakie dokładnie kryterium α miałoby być właściwe w przypadku badań eksploracyjnych. Nie ma i chyba nie powinno być takiej reguły. W świetle rozumowania przedstawionego w poprzednim ustępie jest jednak jasne, że scenariusz eksploracyjny wymaga bardziej rygorystycznych kryteriów istotności statystycznej niż scenariusz konfirmacyjny[7].

---

[5] H_0 oznacza zdarzenie losowe polegające na tym, że testowany parametr populacji przyjmuje wartość zakładaną przez hipotezę zerową. Badany efekt w populacji nie istnieje, zatem ten, który zaobserwowano w próbie, ma charakter nic nieznaczącej, losowej fluktuacji. D oznacza w tym przypadku zaobserwowanie danych, dla których test statystyczny daje wynik $p \leq \alpha$. Zapis w postaci $P(A|B)$ oznacza warunkowe prawdopodobieństwo zajścia zdarzenia A, jeśli zaszło także zdarzenie B.

[6] Większość zależności psychofizycznych ma charakter nieliniowy. Także z gradacją subiektywnej pewności dowodu zdaje się lepiej korespondować szereg prawdopodobieństw błędu 1/10, 1/100, 1/1000 niż 1/10, 1/20, 1/40.

[7] David Colquhoun szacuje procent fałszywych odkryć w badaniach stosujących kryterium $\alpha = 0{,}05$ na około 30%. Przy niewystarczającej mocy statystycznej – dużo więcej [Colquhoun, 2014].

Weryfikacja

Inaczej niż scenariusze opisane wyżej, scenariusz weryfikacyjny pozwala zarówno na odrzucenie, jak i przyjęcie hipotezy zerowej. Daje też wyższą pewność wniosku o populacji. Wszystko to dzięki zrealizowaniu najważniejszego postulatu Neymana i Pearsona, tj. minimalizowaniu nie tylko prawdopodobieństwa fałszywego alarmu (błędu pierwszego rodzaju, α), ale także ryzyka przeoczenia istniejącego efektu (błędu drugiego rodzaju, β).

Typowe badania psychologiczne charakteryzują się małą (często uważa się, że zbyt małą [Cohen, 1992; Sedlmeier i Gigerenzer, 1989; Wilkinson, APA Task Force on Statistical Inference, 1999] mocą statystyczną, $1 - \beta$, a więc także wysokim ryzykiem błędu drugiego rodzaju. Negatywnego wyniku nie sposób użyć jako argumentu na rzecz fałszywości hipotezy zerowej w badaniach, których moc statystyczna $1 - \beta$ wynosi 0,5. Wielokrotnie powtórzone, typowe badanie w naszej dziedzinie byłoby w stanie wykryć faktycznie istniejący efekt tylko w co drugim przypadku. Równie często przyniosłoby wynik fałszywie negatywny. Tę samą częstość trafień uzyska ktoś, kto zamiast kłopotać się badaniami decydowałby o wniosku na podstawie wyników rzutu monetą! To oczywiście uwaga nieco demagogiczna, bo wiarygodność serii niezależnie replikowanych badań jest bez porównania wyższa, trzeba jednak przyznać, że porównanie przemawia do wyobraźni.

Przy wysokiej mocy statystycznej rzecz wygląda zgoła inaczej. Negatywny wynik badań o mocy na przykład $1 - \beta = 0{,}999$ jest bardzo poważnym argumentem przeciw hipotezie obecności efektu w populacji. By umożliwiać nie tylko falsyfikowanie, ale także potwierdzanie hipotezy zerowej, badania prowadzone w scenariuszu weryfikacyjnym muszą się więc charakteryzować wysoką mocą statystyczną. Owa moc jest – dla danej wielkości efektu w populacji, wariancji oraz testu – proporcjonalna do wielkości próby. Dlatego omawiana kategoria badań wymaga zwykle ponadprzeciętnie dużych prób. Trudność realizacji oraz wysokie koszta sprawiają, że badania takie prowadzi się stosunkowo rzadko – wtedy gdy niezbędna jest wysoka pewność wniosku albo badacz potrzebuje mocnego wsparcia tezy o braku efektu w populacji. Przykładu badań realizowanych w scenariuszu weryfikacyjnym dostarcza na przykład projekt McManusa i współpracowników [2013], którego celem było odkrycie genu odpowiedzialnego za dziedziczenie ręczności lub, w przypadku negatywnego wyniku, falsyfikacja jednogenowych modeli dziedziczenia tej cechy. Badania przyniosły mocny wynik negatywny, a pikanterii dodaje im fakt, że sam McManus jest autorem i wieloletnim propagatorem jednego z dwóch powszechnie znanych, bardzo wpływowych modeli jednogenowych.

Pisząc o mocy statystycznej, autorzy opisanego we wstępie raportu APA używają gry słów: *statistical power does not corrupt* („moc statystyczna nie korumpuje") [Wilkinson, APA Task Force on Statistical Inference, 1999, s. 597], wykorzystującej zręcznie fakt, że słowo *power* oznacza w języku angielskim zarówno władzę, jak i moc. Dla efektu retorycznego autorzy przywołują jednak obraz zbyt wyidealizowany, bowiem (utrzymując konwencję bon motu) każda moc niewłaściwie użyta może być niebezpieczna. Moc statystyczna nie jest wyjątkiem. Scenariusz weryfikacyjny wymaga zachowania szczególnej staranności. Badacz musi pieczołowicie zidentyfikować

i usunąć wszelkie zakłócające czynniki systematyczne, których działanie w dużych projektach ulega zwielokrotnieniu. Oddziaływania, które przy małej próbie utonęłyby w szumie zmienności przypadkowej, w badaniach na dużych próbach, charakteryzujących się korzystniejszym stosunkiem sygnału do szumu, stają się widoczne i prowadzą do widocznego skrzywienia wyników. Konieczna jest ścisła kontrola losowości doboru próby, randomizacja przydziału do warunków eksperymentalnych i grup oraz równoważenie potencjalnie zakłócających wpływów kolejności.

Na ogół badacze podejmujący badania w scenariuszu weryfikacyjnym mają wystarczającą świadomość omawianych zagrożeń, by trafność i rzetelność ich projektów nie budziła zastrzeżeń. Częściej zdarzają się błędy interpretacyjne. W badaniach o dużej mocy próg istotności statystycznej łatwo przekraczają efekty o marginalnej istotności merytorycznej. Wysoce „istotna" może się okazać różnica efektywności manipulacji, poziomu zdolności czy poprawności wykonania, zupełnie nieznacząca z praktycznego punktu widzenia. Zalecenie APA szacowania wielkości efektów i obliczania przedziałów ufności dla efektów kluczowych [Wilkinson, APA Task Force on Statistical Inference, 1999] powinno więc być w przypadku scenariusza weryfikacyjnego respektowane szczególnie starannie.

Wreszcie wypada po raz kolejny przypomnieć, że prawdopodobieństwo $P(D|H_0)$ nie jest tożsame z prawdopodobieństwem $P(H_0|D)$, a więc test istotności nie daje bezpośredniej informacji na temat prawdopodobieństwa prawdziwości testowanych hipotez. Nawet przy bardzo dużej mocy i wyśrubowanych poziomach istotności, typowych dla scenariusza weryfikacyjnego, może tylko służyć jako ich przybliżenie. Owo przybliżenie jest tym lepsze, im bardziej prawdopodobieństwa aprioryczne $P(H_0)$ oraz $P(D)$ są zbliżone. W praktyce wynika z tego przede wszystkim postulat nieufności w stosunku do wyliczonych wyników testu istotności w odniesieniu do tych efektów, których istnienie jest *a priori* bardzo mało prawdopodobne[8].

Rzadsze scenariusze

W nauce rzadko zdarza się sytuacja typowa dla innych dziedzin działalności praktycznej, na przykład gospodarczej, kiedy ryzyko błędów I i II rodzaju jest wysokie, ich koszta podobne, a powstrzymanie się od decyzji niemożliwe. Czasem jednak i badacz może być zainteresowany podjęciem optymalnej decyzji na podstawie skąpego materiału. Na przykład projektując badania, z powodu ograniczeń budżetowych nie może włączyć wszystkich potencjalnie interesujących zmiennych i musi, opierając się na wynikach niewielkiego pilotażu, wybrać te najbardziej obiecujące. Jedną z możliwych strategii jest w takim przypadku obniżenie kryterium istotności do poziomu, przy którym do badania zasadniczego kwalifikuje się tyle zmiennych, ile badacz może uwzględnić. Taki scenariusz minimalizowałby ryzyko błędu I i II rodzaju, z priorytetem dla minimalizacji tego ostatniego, co w omawianej sytuacji wydaje się rozsądną strategią.

[8] Test byłby wiarygodny w odniesieniu do wysoce prawdopodobnej *a priori* hipotezy zerowej tylko w przypadku zaobserwowania istotnie sprzecznych z nią, ale zarazem podobnie wysoce prawdopodobnych *a priori* danych, co wydaje się sytuacją czysto teoretyczną.

Istotą weryfikacji hipotez statystycznych jest badanie, na ile sprzeczna z danymi jest hipoteza zerowa zakładająca, że testowany parametr przyjmuje w populacji określoną wartość. Z punktu widzenia logiki samego wnioskowania statystycznego nie ma znaczenia, jaką wartość parametru populacji zakłada hipoteza zerowa. To z punktu widzenia logiki postępowania *badawczego* zwykle najbardziej pożyteczna jest informacja, że z określoną ufnością testowany efekt różni się akurat od zera. Dlatego takie właśnie hipotezy testuje się najczęściej. Najpopularniejsze pakiety statystyczne w podstawowych opcjach nawet nie proponują użytkownikowi innych możliwości. Bywają jednak sytuacje, w których bardziej przydatna byłaby weryfikacja inaczej sformułowanej hipotezy. Na przykład badacza może bardziej interesować nie to, czy różnica efektywności dwóch programów szkoleniowych, mierzona różnicą średnich wyników absolwentów w egzaminie końcowym, jest *różna od zera*, ale czy jest *znacząco duża*, na przykład wynosi więcej niż 10 punktów. Może wtedy, w jednostronnym teście istotności, zbadać, czy i w jakiej mierze dane są sprzeczne z hipotezą zerową, zakładającą dziesięciopunktową przewagę jednego z programów. Jeśli dla założonego poziomu istotności uda mu się odrzucić taką hipotezę, uzyska argument na rzecz istnienia znaczącej (ponaddziesięciopunktowej) przewagi owego programu. Oczywiście taka strategia ma sens tylko wtedy, gdy sytuacja uzasadnia hipotezę kierunkową, czyli w odniesieniu do podanego przykładu, kiedy zasadne jest założenie, że interesujący badacza program *B* jest co najmniej tak skuteczny jak program odniesienia, *A*. W przeciwnym wypadku właściwsze jest obliczenie przedziału ufności dla różnicy.

Weinberg [1979, s. 72] zwracał uwagę, że stosując metody heurystyczne, lepiej wiemy, kiedy ich użyć, niż kiedy tego nie robić. Analogiczną myśl, w odniesieniu do testowania hipotez statystycznych, wyraził Cohen [2006], ironicznie tytułując swój artykuł: „Ziemia jest okrągła ($p < 0{,}05$)". Zdarza się obserwować w badaniach efekty oczywiste, których istotności testować nie ma potrzeby, na przykład dłuższy czas reakcji złożonej niż prostej, mniejszą sprawność pamięci seniorów niż studentów, podwyższony poziom niepokoju u pacjentów lękowych, czy podobne. Czasem badacz przeprowadza test istotności takiego efektu po to, by zademonstrować skuteczność manipulacji eksperymentalnej czy prawidłowość doboru grupy. Nie jest to jednak właściwa droga. Konstatacja: „skuteczność manipulacji została potwierdzona obserwacją wysoce istotnej różnicy między warunkami ($p < 0{,}01$)" jest w istocie innym sposobem powiedzenia: „efekt manipulacji był z dużą dozą pewności różny od zera". To za mało, trzeba pokazać, że ów efekt nie tylko jest większy od zera, ale że jest od niego wystarczająco większy. Można to uzyskać, odrzucając w teście jednostronnym hipotezę zerową zakładającą niedostateczną (nieosiągającą pewnego merytorycznie uzasadnionego minimum) wielkość różnicy. Zwykle jednak łatwiej sprawdzić, czy dolna granica przedziału ufności dla owej różnicy jest wystarczająco wysoka – na przykład czy manipulacja rodzajem bodźca przyniosła wydłużenie czasu reakcji nie mniejsze niż minimalna przewidywana teoretycznie wartość, dajmy na to 200 ms. Jeśli nie ma podstaw do takich oczekiwań, można odwołać się do jakiegoś kryterium ogólnego, na przykład sprawdzić, czy *d* Cohena przekracza umowną granicę efektów dużych, tj. 0,8.

Nie ma sensu testowanie istotności efektów w sposób oczywisty nieprzypadkowych. A co z efektami, które są równie jaskrawo nieznaczące? Spotkałem się niedaw-

no z sytuacją, w której początkujący badacz, testując całą serię mierzonych w badaniach efektów, obliczył z rozpędu także wartość statystyki testowej dla takiego, który miał w próbie wartość równą zero. Wartość p dla tego przypadku można łatwo podać bez testu istotności: prawdopodobieństwo pobrania z dowolnej populacji próby, w której różnica będzie równa lub większa od zera, jest oczywiście równe jedności (zapis $p > 0{,}05$ jest w takim przypadku grubym eufemizmem). Gdy efekt zaobserwowany w próbie wynosi zero, test istotności nie wnosi nic. Zauważmy jednak, że jest podobnie bezużyteczny, gdy ów efekt jest bardzo mały. Meehl [1978] zwracał uwagę, że hipoteza zerowa jest praktycznie zawsze fałszywa, bowiem badane efekty rzadko bywają w populacji zerowe w ścisłym sensie matematycznym. Można powiedzieć obrazowo, że w odróżnieniu od zera arytmetycznego, „zero praktyczne" posiada pewną szerokość: w postępowaniu badawczym częściej niż to, *czy* efekt przekracza zero, liczy się to, *o ile* je przekracza. Jeśli efekt w próbie nie wykracza poza przedział wartości definiujących granice minimalnego znaczenia praktycznego, nawet najmocniejszy test istotności nie wesprze tezy, że przekracza je w populacji.

Estymacja przedziałowa

Estymacja nie należy do scenariuszy testowania istotności, lecz do klasy szerszej – scenariuszy wnioskowania statystycznego. Każdy test istotności można jednak bez straty zastąpić obliczeniem przedziału ufności, dlatego przy wyborze scenariusza testowania istotności należy rozważyć także i tę możliwość. Jeśli dany przedział z ufnością u nie zawiera wartości parametru, określonej przez hipotezę zerową, wówczas ową hipotezę można odrzucić na poziomie istotności $p < 1 - u$. To oczywiste: ktoś, kto ma podstawy, by z 99-procentową ufnością oczekiwać podwyżki między 200 a 300 zł, może być jeszcze bardziej pewien, że podwyżka nie wyniesie zero. Z praktycznego punktu widzenia można więc o przedziale ufności myśleć jak o poszerzonym teście istotności – teście, który oprócz informacji o sprzeczności z hipotezą zerową dostarcza też informacji o wielkości efektu i precyzji szacunku. Przedziały ufności są dlatego coraz częściej rekomendowane jako preferowana forma wnioskowania statystycznego [Brandstätter, 1999; Cumming, 2014].

Szerokość przedziału ufności zmniejsza się, a tym samym dokładność oceny parametru populacji rośnie, wraz z wielkością próby. Estymacja przedziałowa jest więc najbardziej precyzyjna w przypadku dużych prób. Jej użyteczność nie ogranicza się jednak do tych sytuacji. Szerokość przedziału ufności dostarcza bowiem ważnej informacji zwrotnej o rzetelności przeprowadzonych badań, co ma szczególne znaczenie tam, gdzie moc jest niewysoka. W tych przypadkach bardzo wyraźna jest też przewaga przedziałów ufności nad testami istotności: badacz dowiadujący się, że zmierzona korelacja wynosi 0,47 i jest istotna statystycznie na poziomie $p < 0{,}01$, łatwiej ulega złudzeniu posiadania mocnego wyniku niż ten, który wie, że 99-procentowy przedział ufności dla owej korelacji rozciąga się między 0,03 a 0,91, co w praktyce uzasadnia tylko wniosek, że badana zależność nie jest wyraźnie ujemna.

Estymacja przedziałowa jest bardziej niż test istotności odporna na błędne interpretacje [Hoekstra, Johnson i Kiers, 2012]. Nie dotyczą jej dwa największe problemy interpretacyjne testów istotności – utożsamianie istotności statystycznej z praktyczną

oraz uznawanie wyników nieistotnych za potwierdzenie hipotezy zerowej. Odpowiednik trzeciego problemu – złudzenia, że wartość *p* opisuje prawdopodobieństwo istnienia efektu w populacji – dotyczy jednak także estymacji przedziałowej. Użytkownicy często mylnie uważają, że poziom ufności *u* odpowiada prawdopodobieństwu, że szacowany parametr populacji – na przykład średnia albo współczynnik korelacji – znajduje się w wyznaczonym przedziale [Coulson i in., 2010; Hoekstra i in., 2014]. Wartość *u* nie odnosi się jednak do rzetelności jednostkowego przedziałowego szacunku parametru populacji, ale do długofalowej rzetelności *procedury* sporządzania tego szacunku – podaje oczekiwaną częstość, z jaką szacowany parametr populacji mieściłby się w przedziałach ufności, gdyby je wyznaczyć dla wszystkich, możliwych do pobrania z tej populacji, prób prostych.

Opisane wyżej scenariusze różnią się powszechnością. Ze względu na pracochłonność i koszta scenariusz weryfikacyjny jest mimo swoich zalet rzadko używany. Za to w scenariusz konfirmacyjny wpisuje się większość publikowanych badań. Eksploracyjne stosowanie testów istotności jest wprawdzie częste, ale już potrzebne w scenariuszu eksploracyjnym zaostrzenie kryterium istotności nie jest regułą. W badaniach obrazowych mózgu standardowo używa się wyliczanej z testu *t* wartości *p* jako kryterium odróżniania sygnału od szumu, co jest formą selekcyjnego użycia testu istotności. Nie odpowiada ona jednak w pełni opisanemu wcześniej scenariuszowi, bowiem selekcja ta nie ma charakteru wstępnego i nie minimalizuje ryzyka błędu drugiego rodzaju kosztem zwiększenia prawdopodobieństwa fałszywego alarmu. Ten ostatni element jest obecny w badaniach psychologicznych nie wprost, pod postacią praktyki wyróżniania oprócz efektów istotnych także klasy efektów granicznych (np. $p = 0{,}071$), nazywanych mylnie „trendami" lub „tendencjami". Mimo iż popularna, praktyka ta nie jest właściwa, gdyż w odniesieniu do jednostkowego efektu trudno mówić o trendzie czy tendencji w ścisłym znaczeniu kierunku zmian. W języku potocznym używa się tych terminów również na oznaczenie subiektywnych predylekcji, tu jednak nie chodzi o to, że efekt nie przekroczył progu istotności, ale miał na to ochotę. Podręcznik APA [American Psychological Association, 2010] potępia omawiany zwyczaj jako niezgodny z zasadami wnioskowania statystycznego, lecz nie podsuwa naturalnego rozwiązania problemu w postaci uwolnienia kryterium *alfa*. Zamiast ślepo stosować zawsze to samo konwencjonalne kryterium $\alpha = 0{,}05$, badacz powinien korzystać z wolności wyboru kryterium, odzwierciedlającego jego *prawdziwe* wyobrażenie o tym, jaki poziom istotności jest właściwy w danym zastosowaniu. I tak w scenariuszu selekcyjnym – zamiast gimnastykować się z naciągającymi logikę wnioskowania „trendami" – powinien użyć liberalnego kryterium istotności, na przykład $\alpha = 0{,}1$. Oczywiście pod warunkiem konsekwentnego respektowania prowizoryczności uzyskanego w ten sposób słabego wniosku.

Naturalną konsekwencją rozumienia istotności statystycznej jako *wystarczającej sprzeczności* z hipotezą zerową wydaje się uznanie, że różne okoliczności wymagają różnych poziomów owej sprzeczności i odrzucenie praktyki stosowania jednego, sztywnego kryterium. Jak zatem decydować o istotności testowanych efektów? Fisher pisał: „Mimo wygody, jaką daje konstatowanie sprzeczności z hipotezą [zerową] na pewnym typowym poziomie istotności, takim jak 5%, 2% czy 1%, nie powinniśmy we wnioskowaniu indukcyjnym nigdy tracić z oczu dokładnej wagi uzyskanego dowodu

ani zapominać, że w toku dalszych badań może się ona zwiększyć lub zmniejszyć”[9]. W tej dość dyplomatycznej wypowiedzi statystyk z jednej strony podkreśla potrzebę indywidualnej oceny wartości *p*, a z drugiej docenia wartość ogólniejszych standardów. Praktyka badań psychologicznych w istocie zbliża się do tego zalecenia Fishera, bowiem wartość α = 0,05 traktuje się jako *de facto* standard minimalny – czyniąc wyjątek dla wspomnianych „trendów” (*sic!*), do których stosuje się liberalne minimum w granicach 0,07–0,1. W przypadku gdy wartość *p* przekracza próg 0,05 z dużym zapasem, sygnalizuje się to, oznaczając efekt w tabelach i na wykresach gwiazdkami symbolizującymi kolejne progi istotności, najczęściej: 0,01 i 0,001. APA zaleca od pewnego czasu także rutynowe podawanie dokładnej wartości *p* [Wilkinson, APA Task Force on Statistical Inference, 1999]. Mamy więc do czynienia ze współistnieniem powszechnie uznawanych standardów (kryteria α) oraz indywidualnej oceny (jakościowa interpretacja wartości *p*). Te pierwsze, dla uniknięcia życzeniowej interpretacji, powinny być i są przyjmowane przed zrealizowaniem badań, *a priori* i traktowane jako, różne dla różnych scenariuszy, standardy minimalne. Natomiast dokonywana *a posteriori* ocena wiarygodności wyniku na podstawie wartości *p* pozwala wyjść poza samą konstatację spełnienia minimalnego standardu istotności i ocenić faktycznie uzyskany – często znacznie wyższy – poziom wiarygodności poczynionej obserwacji.

Podsumowanie

W niniejszym cyklu artykułów poświęconych praktyce testowania istotności w psychologii moim głównym celem było zachęcenie do zastąpienia niewoli rytuału samodzielną refleksją – myśleniem statystycznym i metodologicznym, którego istotą jest świadome budowanie narracji opisującej i uzasadniającej konstruowany model opisywanego fragmentu rzeczywistości. Mimo że siłą rzeczy początkujący badacze koncentrują się bardziej na pytaniu *jak?*, nie powinni jednak zaniedbywać jeszcze ważniejszego pytania *po co?*

Istotność statystyczna znaczy dużo mniej, niż się powszechnie sądzi. Niezrozumienie tej prawdy skutkuje praktyką rytualnego polowania na efekty spełniające zaklęty warunek $p \leq 0{,}05$, nazywane czasem celnie *p-hacking* [Simonsohn, Nelson i Simmons, 2014]. Upowszechnianie rozumienia właściwego sensu istotności można zacząć od zastąpienia, tam gdzie to możliwe, mylącej i niejasnej etykiety „istotny statystycznie” bardziej precyzyjnym, bliższym fisherowskich korzeni sformułowaniem „znacząco sprzeczny z hipotezą zerową”. W odniesieniu do większości typowych sytuacji taka konstatacja znaczy tylko, że analizowany efekt jest w populacji prawdopodobnie różny od zera, a to nie kończy, lecz rozpoczyna poznawanie badanej zależności.

[9] „Convenient as it is to note that a hypothesis is contradicted at some familiar level of significance such as 5% or 2% or 1% we do not, in Inductive Inference, ever need to lose sight of the exact strength which the evidence has in fact reached, or to ignore the fact that with further trial it might come to be stronger, or weaker” [Fisher, 1971, s. 25].

BIBLIOGRAFIA

Abelson, R.P. (1995). *Statistics as Principled Argument*. New York: Psychology Press. Taylor & Francis Group.

American Psychological Association. (2010). *Publication Manual of the American Psychological Association. 6th edition*. Washington, DC: American Psychological Association.

Brandstätter, E. (1999). Confidence intervals as an alternative to significance testing. *Methods of Psychological Research Online*, *4*(2), 33–46.

Brzeziński, J.M. (2012a). Co to znaczy, że wyniki przeprowadzonych przez psychologów badań naukowych poddawane są analizie statystycznej? *Roczniki Psychologiczne*, *15*(3), 7–40.

Brzeziński, J.M. (2012b). Kontekst teorii psychologicznej a kontekst analizy statystycznej. *Roczniki Psychologiczne*, *15*(3), 75–81.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159.

Cohen, J. (1994). The earth is round (p < .05). *American Psychologist*, *49*(12), 997. doi:10.1037/0003-066X.49.12.997

Cohen, J. (2006). Ziemia jest okrągła (p < 0,05). W: J. Brzeziński, J. Siuta (red.), *Metodologiczne i statystyczne problemy psychologii* (s. 100–118). Poznań: Wydawnictwo Zysk i S-ka.

Colquhoun, D. (2014). An investigation of the false discovery rate and the misinterpretation of p-values. *Royal Society Open Science*, *1*(3), 140216. doi:10.1098/rsos.140216

Coulson, M., Healey, M., Fidler, F., Cumming, G. (2010). Confidence intervals permit, but do not guarantee, better inference than statistical significance testing. *Frontiers in Psychology*, *1*, 26. doi:10.3389/fpsyg.2010.00026

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*(1), 7–29. doi:10.1177/0956797613504966

Finch, S., Cumming, G., Williams, J., Palmer, L., Griffith, E., Alders, C., ... Goodman, O. (2004). Reform of statistical inference in psychology: The case of *Memory & Cognition*. *Behavior Research Methods, Instruments & Computers*, *36*(2), 312–324.

Fisher, R.A. (1971). *The Design of Experiments* (wyd. 8). New York: Hafner Publishing Company.

Gigerenzer, G. (1998). We need statistical thinking, not statistical rituals. *Behavioral and Brain Sciences*, *21*, 199–200.

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

Greene, J., D'Oliveira, M. (1982). *Open Guides to Psychology: Learning to Use Statistical Tests in Psychology: A Student's Guide*. Milton Keynes: Open University Press.

Hoekstra, R., Johnson, A., Kiers, H.A.L. (2012). Confidence intervals make a difference: Effects of showing confidence intervals on inferential reasoning. *Educational and Psychological Measurement*, *72*(6), 1039–1052. doi:10.1177/0013164412450297

Hoekstra, R., Morey, R.D., Rouder, J.N., Wagenmakers, E.J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin and Review*, *21*(5), 1157–1164. doi:10.3758/s13423-013-0572-3

Kahneman, D. (2011). *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux.

McManus, I.C., Davison, A., Armour, J.A. (2013). Multilocus genetic models of handedness closely resemble single-locus models in explaining family data and are compatible with genome-wide association studies. *Annals of the New York Academy of Sciences*, *1288*, 48–58. doi:10.1111/nyas.12102

Meehl, P.E. (1978). Theoretical risks and tabular asterisks: Sir Karl, sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*(4), 806–834.

Ostasiewicz, W. (2012). *Myślenie statystyczne*. Warszawa: Oficyna a Wolters Kluwer Business.

Palij, M. (2012). New statistical rituals for old. *PsycCRITIQUES*, *57*(24). doi:10.1037/a0028079

Sedlmeier, P., Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, *105*(2), 309.

Simonsohn, U., Nelson, L.D., Simmons, J.P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, *143*(2), 534–547. doi:10.1037/a0033242

StatsLife. (2015). Academic journal bans p-value significance test. Royal Statistical Society. Pobrane z: http://www.statslife.org.uk/news/2116-academic-journal-bans-p-value-significance-test.

Trafimow, D., Marks, M. (2015). Editorial. *Basic and Applied Social Psychology*, *37*(1), 1–2. doi:10.1080/01973533.2015.1012991

Tversky, A., Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, *76*(2), 105.

Wasserstein, R. (2015). ASA comment on a journal's ban on null hypothesis statistical testing. Pobrane z: http://community.amstat.org/blogs/ronald-wasserstein/2015/02/26/asa-comment-on-a-journals-ban-on-null-hypothesis-statistical-testing.

Weinberg, G.M. (1979). *Myślenie systemowe*. Warszawa: Wydawnictwa Naukowo-Techniczne.

Wilkinson, L., APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*(8), 594–604.

Wojciszke, B. (2004). Systematycznie modyfikowane autoreplikacje: Logika programu badań empirycznych w psychologii. W: J. Brzeziński (red.), *Metodologia badań psychologicznych. Wybór tekstów* (s. 44–60). Warszawa: Wydawnictwo Naukowe PWN.

Woolston, Ch. (2015). Psychology journal bans P values. *Nature*, *519*(7541), 9–9. doi:10.1038/519009f