Rocznik Kognitywistyczny 9/2016, s. 59–70
doi: 10.4467/20843895RK.16.006.6412
www.ejournals.eu/Rocznik-Kognitywistyczny

PIOTR WOLSKI

Uniwersytet Jagielloński
Instytut Psychologii

Istotność statystyczna II. Pułapki interpretacyjne

Statistical significance II. Interpretive pitfalls

Abstract: The second of the series of essays on the problems of significance testing in psychological research focuses on inconsistencies of the logic of these tests and resulting problems with interpretation. The limits of their practical usability have been discussed, and reasons of their failure with a priori unlikely null-hypotheses explained. Misleading connotations of the term "statistical significance" have been stressed, that obscure the true meaning of statistical significance and promote bad practices, including overestimation of significance, and neglecting the problem of effect size.

Keywords: statistical inference, null hypothesis significance testing, NHST, p-value, reversed conditional error

Bardzo wiele niedostatków analiz statystycznych w pracach psychologicznych wynika, mniej lub bardziej bezpośrednio, z tego samego podstawowego problemu – przeceniania znaczenia źle rozumianej istotności statystycznej. Problem nabiera tym większej wagi, że – jak pokazują dane omówione w pierwszej części – niewłaściwa interpretacja testów istotności jest alarmująco powszechna. Psychologowie łatwo wpadają w poczucie winy z powodu nie dość starannego odrobienia lekcji statystyki, jednak na przeszkodzie efektywnej dydaktyki wnioskowania statystycznego stają także inne czynniki, których omówienie zawiera niniejsza, druga część cyklu poświęconego problemom praktycznej aplikacji testów istotności w badaniach naukowych w psychologii.

Życzeniowa interpretacja *p*

Przypomnijmy, że *p* to teoretyczne prawdopodobieństwo wystąpienia efektu równego zaobserwowanemu w badaniach lub większego, w próbie pobranej z hipotetycznej populacji o zerowej wartości efektu[1], zaś α to poziom istotności, czyli arbitralne kryte-

[1] O ile definiowana hipotezą zerową wartość *parametru* populacji nie musi wynosić zero, o tyle związany z nią w realnych badaniach *efekt* jest praktycznie zawsze zerowy.

rium określające akceptowane przez badacza ryzyko fałszywego alarmu, zwyczajowo najczęściej ustalane na 0,05.

W części pierwszej zwracaliśmy uwagę na różnicę między prawdopodobieństwem odrzucenia hipotezy zerowej, *gdy* jest ona prawdziwa, a prawdopodobieństwem, *że* odrzucona hipoteza zerowa (H_0) jest prawdziwa. Różnicę między tymi dwoma prawdopodobieństwami widać jasno, gdy się je zapisze w postaci warunkowej[2]. Prawdopodobieństwo fałszywie pozytywnego wyniku testu istotności to $P[(p \leq \alpha)|H_0]$, a prawdopodobieństwo, że mimo pozytywnego wyniku testu hipoteza zerowa jest prawdziwa, to $P[H_0|(p \leq \alpha)]$. Zaobserwowanie danych empirycznych, dla których wylicza się statystykę testową i sprawdza warunek $p \leq \alpha$, oznacza się często jako D, co pozwala zapisać kluczowe tutaj prawdopodobieństwa ogólniej, odpowiednio: $P(D|H_0)$ i $P(H_0|D)$.

Ta ostatnia wielkość wymaga komentarza: o ile prawdopodobieństwo zaobserwowania określonej wartości statystyki w próbie pobranej z populacji o danej wartości parametru, $P(D|H_0)$, jest wielkością, której interpretacja nie nastręcza trudności, o tyle prawdopodobieństwo odwrotne – posiadania przez parametr populacji określonej wartości, przy danej wartości statystyki w próbie, $P(H_0|D)$ – jest wielkością o niejasnym znaczeniu. W kontekście klasycznego, fisherowskiego wnioskowania statystycznego nie ma sensu, bowiem wartość parametru populacji nie jest w tym ujęciu traktowana jak zmienna losowa, lecz jak nieznana stała. Dowolna hipoteza na temat jej wartości jest w chwili sformułowania – obiektywnie rzecz biorąc – prawdziwa lub fałszywa, bez stanów pośrednich. Stopniować można tylko *niepewność badacza* co do prawdziwości lub fałszu hipotezy, co jednak wymaga odwołania się do subiektywnego rozumienia prawdopodobieństwa. Takie rozumienie, dopuszczalne w podejściu bayesowskim, jest odrzucane przez zwolenników klasycznego podejścia frekwencyjnego [Fisher, 1971]. Pozostawiając statystykom spory co do właściwego rozumienia prawdopodobieństwa, w przedstawionych tu i dalej rozważaniach przyjmuję perspektywę badacza praktyka, dla którego sprawą ważniejszą od formalnej czystości testu statystycznego jest jego efektywność w roli narzędzia wspomagającego postępowanie badawcze.

Wróćmy do kwestii rozumienia wartości p. Badacz testujący hipotezę H_0 gromadzi dane D i chciałby wiedzieć, jakie jest w ich świetle prawdopodobieństwo, że owa hipoteza jest prawdziwa, $P(H_0|D)$, bądź fałszywa, $P(H_1|D) = 1 - P(H_0|D)$ [por. Gigerenzer, 2004]. Chciałby się na przykład dowiedzieć, że skoro dla badanego efektu statystyka t uzyskała wartość 2,5, to przy danej liczbie stopni swobody prawdopodobieństwo, iż hipoteza zerowa jest prawdziwa, wynosi 0,02 – a tym samym prawdopodobieństwo trafności hipotezy alternatywnej równa się: 1 − 0,02 = 0,98. Test istotności nie daje jednak takich informacji, a badacz musi się zadowolić czymś innym: dowiaduje się, jakie *byłoby* prawdopodobieństwo wystąpienia efektu o wielkości równej lub większej od faktycznie zaobserwowanego, *gdyby hipoteza* H_0 *była prawdziwa*. Zamiast $P(H_0|D)$ poznaje więc prawdopodobieństwo $P(D|H_0)$, oznaczane w testach istotności literą P lub p. Jeśli p jest małe, przyjmuje, że także hipoteza zerowa jest mało prawdopodobna,

[2] Prawdopodobieństwo zdarzenia A pod warunkiem B, zapisywane $P(A|B)$, to – w najbardziej intuicyjnej interpretacji frekwencyjnej – częstość zdarzenia A oczekiwana w sytuacji, w której wiadomo, iż zaszło zdarzenie B.

a zatem można ją odrzucić. Logika tego wniosku jest analogiczna do stosowanej na przykład przez psychologa, który widząc objawy depresji w nasileniu rzadko obserwowanym u osób zdrowych, wnioskuje, że ma do czynienia z chorobą.

Omówione w części pierwszej badania Oakesa [1986] oraz Hallera i Kraussa [2002] pokazują, że przytłaczająca większość studentów, badaczy oraz nauczycieli życzeniowo interpretuje p, czyli $P(\mathrm{D}|H_0)$, jako $P(H_0|D)$. Innymi słowy, traktuje prawdopodobieństwo zaobserwowania określonych danych w przypadku prawdziwości hipotezy zerowej, jakby to było prawdopodobieństwo prawdziwości tej hipotezy w świetle zaobserwowanych danych. To błąd logiczny, znany jako „błąd odwrócenia warunku" (*fallacy of the transposed conditional* [Wagenmakers i in., 2011]). Prawnicy nazywają go „błędem prokuratorskim", bo zwiększa ryzyko niekorzystnej dla oskarżonego interpretacji materiału dowodowego. Wyobraźmy sobie, że nieznany członek społeczności, liczącej 10 tysięcy osób, dokonał zabójstwa, zostawiając na miejscu zbrodni ślady krwi bardzo rzadkiej grupy, występującej u zaledwie 0,1% populacji. Krew podejrzanego należy do tej właśnie grupy. Prokurator wierzy, że to mocny dowód jego winy: uważa, że skoro prawdopodobieństwo zaobserwowania krwi z owej grupy (D) u osoby niewinnej (H_0) wynosi 1/1000, to takie właśnie jest prawdopodobieństwo, że podejrzany jest niewinny. Myli się jednak prawie o trzy rzędy wielkości, bowiem choć pierwsze z tych prawdopodobieństw, $P(D|H_0)$, faktycznie równa się 1/1000, to drugie, $P(H_0|D)$, wynosi 10/11[3]. Nie mamy więc do czynienia z praktyczną pewnością winy, ale przeciwnie – z wysokim, bo aż 91-procentowym prawdopodobieństwem niewinności [Fenton i Neil, 2011]. Także inne przykłady przekonują, że „odwrócone" prawdopodobieństwa warunkowe potrafią się zasadniczo różnić. Z tego, że niewielu mężczyzn jest neurochirurgami, nie wynika wcale, że podobnie niewielu neurochirurgów jest mężczyznami; prawdopodobieństwo przyspieszonego tętna u osoby przeżywającej atak lękowy jest dużo wyższe niż prawdopodobieństwo, że osoba, której tętno jest przyspieszone, przeżywa akurat atak lękowy – bardziej prawdopodobne, że idzie szybkim krokiem albo wchodzi po schodach. Podobnie błędne jest rozumowanie badacza, który z uzyskanej w teście istotności wartości p – czyli $P(D|H_0)$ – równej na przykład 0,012 wnosi, że tyle właśnie wynosi ryzyko, iż testowany efekt „w rzeczywistości" – czyli w populacji – nie istnieje. Faktyczne ryzyko fałszywego alarmu $P(H_0|D)$ może być bowiem dużo większe (albo mniejsze) od p.

Badania pokazują, że mylenie $P(D|H_0)$ z $P(H_0|D)$ jest powszechne. Warto zatem sprawdzić, do jakich konsekwencji arytmetycznych prowadzi ów często popełniany błąd logiczny. $P(D|H_0)$ to p. Oznaczmy dla wygody $P(H_0|D)$ jako h. Z formuły Bayesa wynika, że h jest równe p pomnożonemu przez iloraz apriorycznych prawdopodobieństw hipotezy zerowej i danych: $P(H_0)/P(D)$. Zatem gdy $P(H_0) = P(D)$, wówczas h równa się p i błędu nie ma. Gdy $P(H_0) < P(D)$, wtedy $p > h$, a więc badacz interpretujący p jako h przecenia ryzyko fałszywego alarmu i w konsekwencji ocenia szanse istnienia poszukiwanego efektu ($1 - h$) konserwatywnie. Jeśli $P(H_0) > P(D)$, jest w tej

[3] Wynika to z następującego obliczenia: jedna z 10 tysięcy osób ma charakterystyczną grupę krwi i jest zabójcą; 0,1% z pozostałych 9999 osób, czyli (w przybliżeniu) 10 osób, także ma tę samą, rzadką grupę krwi, ale jest niewinne; skoro więc zabójcą jest jedna z ogółem jedenastu osób o tej grupie krwi, prawdopodobieństwo niewinności każdej z nich, w tym także podejrzanego, wynosi 10/11, czyli około 0,91.

ocenie liberalny. Wspomnieliśmy wyżej o trudnościach interpretacyjnych związanych z wartością warunkowego prawdopodobieństwa hipotezy zerowej, $P(H_0|D)$. Podobnie niejednoznaczna jest kwestia interpretacji prawdopodobieństw apriorycznych $P(H_0)$ oraz $P(D)$, nie mówiąc już o trudności wyznaczenia tych wartości dla konkretnych badań. Jeśli jednak przyjmiemy, że $P(H_0)$ koresponduje z potocznie rozumianymi szansami na to, iż postawiona przez badacza hipoteza zerowa jest prawdziwa, a $P(D)$ ze stopniem, w jakim zaobserwowane dane można uznać za typowe, wówczas systematyczne skrzywienie „odwróconego testu istotności", w którym *p* jest interpretowane jako *h*, zyskuje następujący wyraz, zależny od relacji między prawdopodobieństwami *a priori*: Jeśli uzyskano względnie typowe dane, a istnienie efektu jest *a priori* relatywnie wysoko prawdopodobne, czyli zachodzi nierówność $P(H_0) < P(D)$, odwrócony test istotności działa konserwatywnie i nie docenia rzeczywistego prawdopodobieństwa istnienia efektu w populacji, $P(H_1|D)$; odwrotnie, gdy $P(H_0) > P(D)$ – odwrócony test istotności staje się wówczas liberalny. Interpretując niskie *p* jako świadectwo wysokiego prawdopodobieństwa obecności efektu w populacji, badacz ponosi szczególnie duże ryzyko błędu pierwszego rodzaju (fałszywego alarmu) wtedy, gdy jego hipoteza badawcza jest – w świetle wcześniejszych wyników lub rozważań teoretycznych – mało prawdopodobna albo dane, które zaobserwował, są nietypowe. To ciekawe, że odruchowa nieufność w stosunku do nieoczekiwanych wyników chroni badaczy przed popełnianiem właśnie takich błędów interpretacyjnych. Wygląda na to, że wszyscy jesteśmy po trosze instynktownymi bayesowcami.

Badacz na ogół nie dysponuje danymi, które pozwoliłyby mu wiarygodnie oszacować iloraz $P(H_0)/P(D)$, więc nie wiadomo, o jaki czynnik przeszacowuje lub niedoszacowuje *h*, jeśli ocenia je na podstawie *p*. Mogłoby się wydawać, że nie ma to większego znaczenia, bowiem skoro interpretowanie wartości *p* jako *h* prowadzi do błędu odwróconego warunku, to należy się od takich interpretacji po prostu powstrzymać. Logika odwróconego warunku nie jest jednak tylko błędem interpretacyjnym. Co gorsza, wydaje się także immanentną słabością samego testu istotności, a ściślej – sylogizmu, na którym ten się zasadza [Kalinowski, Fidler i Cumming, 2008; Westover, Westover i Bianchi, 2011].

Dyskusyjny sylogizm

Fisher szukał rozwiązania ważnego paradoksu metodologicznego: jedyny sposób pozyskiwania nowej wiedzy o świecie, wnioskowanie indukcyjne, opiera się na empirycznej weryfikacji teorii; tymczasem statystyczny sposób wnoszenia o przyczynach z obserwacji skutków bazuje na formule Bayesa i związanym z nią pojęciu „prawdopodobieństwa odwrotnego", które Fisher – podobnie jak wielu prominentnych statystyków tamtego okresu – uważał za niewłaściwe. Ujęcie bayesowskie zmuszało bowiem do porzucenia rozumienia prawdopodobieństwa jako obiektywnej wielkości, znajdującej swój obserwowalny wyraz w częstości zdarzeń, na rzecz subiektywnej „zaledwie skłonności psychologicznej" [Fisher, 1971, s. 7], co zdawało się czynić je bezużytecznym w zastosowaniach naukowych. Faktycznie, o ile pytanie na przykład o prawdopodobieństwo pobrania próby o średniej większej od 15 z populacji o średniej równej

zero ma sensowną w kategoriach częstościowych, policzalną odpowiedź, o tyle już odwrócone pytanie o prawdopodobieństwo, że średnia w populacji wynosi zero, jeśli w próbie jest większa niż 15, nie ma przy frekwencyjnym rozumieniu prawdopodobieństwa jasnego sensu. Mowa wszak o – trudnym do rozważania w kategoriach częstości – zdarzeniu jednostkowym, które do tego już się zrealizowało, a więc jego obiektywne prawdopodobieństwo wynosi jeden, jeśli hipoteza zerowa jest prawdziwa, albo zero, jeśli jest fałszywa. Prawdopodobieństwo odwrotne nie mówi o obiektywnym stanie rzeczywistości, tylko o zasadności subiektywnego przekonania o tym stanie, co było dla Fishera nie do przyjęcia. Zamiast o prawdopodobieństwie wolał w tym kontekście mówić o tzw. wiarygodności (*likelihood*). W efekcie opracował metodę wnioskowania indukcyjnego – test istotności – która pozwalała weryfikować przewidywania teoretyczne bez odwoływania się do, intelektualnie wątpliwej, konstrukcji prawdopodobieństwa odwrotnego.

Istota pomysłu była prosta: badacz formułuje hipotezę zerową, określając tym samym nieznaną wielkość efektu w populacji, co przy znanym rozkładzie i wartości błędu standardowego umożliwia obliczenie prawdopodobieństwa wystąpienia danego zakresu wielkości efektów w próbie. Jeśli zaobserwowane w badaniu dane są wyraźnie sprzeczne z przewidywaniami wynikającymi z założenia hipotezy zerowej, badacz odrzuca tę hipotezę, uzyskując w ten sposób pośrednie potwierdzenie hipotezy badawczej. Fisher nawiązał tu do schematu rozumowania *modus tollens*: jeśli następnik implikacji jest fałszywy, to fałszywy jest także poprzednik. Zamiast kanonicznej postaci sylogizmu musiał jednak użyć słabego wariantu probabilistycznego. Mocny sylogizm wyglądałby tak: „Jeśli hipoteza H_0 jest prawdziwa, zaobserwowanie danych D jest *niemożliwe*; zaobserwowano dane D, zatem hipoteza H_0 jest *fałszywa*". Poprawność logiczna tego rozumowania nie budzi wątpliwości. Nie można jednak powiedzieć tego samego o słabej wersji: „Jeśli hipoteza H_0 jest prawdziwa, zaobserwowanie danych D jest *mało prawdopodobne*; zaobserwowano dane D, zatem hipoteza H_0 jest *mało prawdopodobna*". W terminach prawdopodobieństw warunkowych istota tego rozumowania sprowadza się bowiem do dyskusyjnej implikacji: „Jeśli prawdopodobieństwo $P(D|H_0)$ jest małe, to i prawdopodobieństwo $P(H_0|D)$ jest małe". Widzieliśmy wcześniej, że odwrócone prawdopodobieństwa warunkowe są sobie równe tylko wtedy, gdy $P(H_0) = P(D)$. Zatem w takiej mierze, w jakiej owo założenie nie jest spełnione, błąd odwróconego warunku dotyczy wszystkich użytkowników testów istotności, a nie tylko tych, którzy błędnie interpretują wartość p jako prawdopodobieństwo $P(H_0|D)$, czyli h[4].

[4] Fisher wątpił w zasadność bayesowskiego „prawdopodobieństwa odwrotnego", więc nie uznałby odwołującego się doń zarzutu. Protestowałby też przeciwko utożsamieniu decyzji o roboczym odrzuceniu hipotezy z orzeczeniem jej niskiego prawdopodobieństwa. Trudno jednak zaprzeczyć, że podstawą owej decyzji jest większa zasadność oczekiwania pozytywnej niż negatywnej replikacji, a to może uzasadniać mówienie o prawdopodobieństwie, nawet w klasycznym rozumieniu frekwencyjnym. Kwestionując subiektywne rozumienie prawdopodobieństwa, Fisher wskazywał na brak powszechnej akceptacji formuły Bayesa oraz jej bardzo nieliczne, zważywszy na powszechną znajomość, aplikacje [Fisher, 1971, s. 6–7]. Dzisiejsza popularność statystyki bayesowskiej zdecydowanie osłabia ten argument. Także spór między zwolennikami częstościowego i subiektywnego rozumienia prawdopodobieństwa nie ma już dawnej ostroś-

Powyższa konstatacja ma ważne implikacje praktyczne. Użyteczność testu istotności warunkuje bowiem ta sama reguła, którą opisaliśmy wyżej, analizując sytuację, w której badacz interpretuje błędnie prawdopodobieństwo p jako h: wniosek na temat istotności jest wiarygodny o tyle, o ile aprioryczne prawdopodobieństwo hipotezy zerowej jest zbliżone do apriorycznego prawdopodobieństwa zaobserwowanych danych. Przykładowo, dla danych, dla których $P(D) = 0{,}5$, test jest konserwatywny, gdy hipoteza alternatywna jest *a priori* bardziej prawdopodobna od zerowej, a liberalny w przeciwnym przypadku, gdy bardziej prawdopodobna jest hipoteza zerowa. Skrzywienie testu jest proporcjonalne do wielkości odchyłki od wartości 0,5.

Badacz rzadko potrafi dokładnie oszacować aprioryczne prawdopodobieństwa $P(H_0)$ i $P(D)$, ale też rzadko zdarza się, by nie wiedział o nich zupełnie nic. Zidentyfikowanie skrajnych przypadków zwykle nie jest trudne. Z czterech skrajności, $P(D) \approx 0$, $P(D) \approx 1$, $P(H_0) \approx 0$ oraz $P(H_0) \approx 1$, największe znaczenie praktyczne ma ta ostatnia. Wyniki skrajnie nieoczekiwane $P(D) \approx 0$ nie są akceptowane bez dodatkowej starannej weryfikacji (replikacji), a poza tym – z definicji – prawie się nie zdarzają. Gdy $P(D) \approx 1$, niepewność co do spodziewanego wyniku jest tak mała, że raczej nie szuka się rozstrzygnięć empirycznych. Nie szuka się ich też, mając praktyczną pewność fałszu hipotezy zerowej $P(H_0) \approx 0$. Co innego, gdy $P(H_0) \approx 1$. Badacze próbują czasem uwiarygodnić za pomocą testu istotności hipotezy, które są *a priori* wyjątkowo mało prawdopodobne. Przychodzi tu na myśl kontrowersyjny przypadek ogłoszonych przez Daryla Bema [Bem, 2011; Bem, Utts i Johnson, 2011] rzekomych dowodów prekognicji [zob. krytyka w Wagenmakers i in., 2011]. W takich sytuacjach zachodzi nierówność $P(H_0) \gg P(D)$, a test istotności staje się liberalny w stopniu praktycznie negującym jego użyteczność. Jak w opisanym wcześniej przypadku „błędu prokuratorskiego", odwrócenie warunku prowadzi do radykalnego niedoszacowania ryzyka fałszywego alarmu.

Zauważmy, że powyższe rozważania stawiają w innym świetle, opisywane w pierwszej części, wyniki Oakesa [1986] oraz Hallera i Kraussa [2002]. Jeśli bowiem poprawność słabego sylogizmu probabilistycznego, a wraz z nią wiarygodność testu istotności, opiera się na założeniu praktycznej ekwiwalentności p i h, trudno mieć za złe użytkownikom, że powszechnie traktują te dwa prawdopodobieństwa jako tożsame…

Jestem daleki od sugerowania, że interpretowanie p jako h jest właściwe. Myślę jednak, że w praktyce ów popularny błąd wyrządza efektywności procesu badawczego szkodę mniejszą niż dwa inne błędy w interpretacji istotności. Pierwszy to rzadko wyrażane jawnie, ale częste *implicite*, traktowanie pozytywnego wyniku testu jako potwierdzenia zgodności danych z próby z danymi z populacji. Popełniający ten błąd badacz ma skłonność do nieuzasadnionego przyjmowania, że jeśli na przykład grupa E uzyskała w badaniach wynik o 18 punktów większy od grupy K, a $p \leq 0{,}05$, to znaczy, że – z marginesem niepewności proporcjonalnym do wartości p – wyniki grupy E są „naprawdę" (czyli w populacji) wyższe o około 18 punktów od wyników grupy K. Ta forma estymacji nie ma mocnych podstaw statystycznych[5], ale wydaje się uza-

ci, odkąd dominującą pozycję zyskała w statystyce definicja aksjomatyczna, Kołmogorowa [Westover, Westover, Bianchi, 2011].

[5] Wprawdzie średnia z próby faktycznie jest najlepszym estymatorem średniej w populacji, jednak określenie poziomu istotności/ufności wymaga zastosowania estymacji przedziałowej, a nie punktowej.

sadniona psychologicznie: ponieważ test istotności jest życzeniowo traktowany jako narzędzie statystycznej walidacji badań, spełnienie warunku $p \leq \alpha$ wydaje się stanowić swoisty certyfikat wiarygodności – *nihil obstat* dla generalizacji wyników. Badacz ulegający omawianej tendencji interpretacyjnej nie pyta o wielkość efektu w populacji, bo ma wrażenie, że ją zna: spodziewa się, że jest podobna jak w próbie. Dlatego nawet jeśli, kierując się zaleceniami APA, oblicza przedziały ufności, to pozostawia je bez interpretacji, gdyż ich w swoim mniemaniu nie potrzebuje. Drugi – powiązany z powyższym – błąd, przed którym ostrzegał już Fisher, to mylenie istotności statystycznej z istotnością praktyczną [Kirk, 1996].

Mylące konotacje terminu „istotność statystyczna"

Pojęcie istotności statystycznej jest w psychologii przywoływane bardzo często. Badacze piszą o statystycznie istotnych różnicach, korelacjach, wynikach, oddziaływaniach, zmianach, słowem – efektach. Fisher używał tego terminu inaczej – zamiast o istotnych statystycznie efektach mówił raczej o ich istotnej lub znaczącej *niezgodności z hipotezą zerową* [np. Fisher, 1971, s. 15]. Z technicznego punktu widzenia oba sformułowania są równoważne, bowiem statystycznie istotny jest z definicji właśnie ten efekt, którego obecność w próbie znacząco kłóci się z hipotezą zerową. Ich znaczenia konotacyjne są jednak zasadniczo odmienne – efekt istotny znaczy w potocznym rozumieniu tyle, co znaczący, ważny, zasługujący na uwagę, natomiast efekt niezgodny z hipotezą zerową to tylko efekt niezerowy, nieznanej wielkości, wymagający dalszej eksploracji – jeśli interesujący, to raczej tylko potencjalnie. Drugi zestaw skojarzeń jest oczywiście bliższy właściwemu sensowi pojęcia: efekt statystycznie istotny to efekt z wystarczającą pewnością niezerowy, nic więcej. Nawet owa „wystarczająca pewność" jest jak na standardy naukowe raczej niska, skoro poziomem odniesienia jest dla niej zwyczajowo ryzyko błędu równe 5%. Laik byłby zdumiony, słysząc, że badacz, który mówi o statystycznie istotnym wzroście wynagrodzeń, ma na myśli tylko tyle, iż ten prawdopodobnie nie wynosi zero. Wydawałoby się przecież, że mowa o wzroście, który nie tylko jest znaczący, ale też dzięki użyciu zaawansowanych metod statystycznych potwierdzono, że dokumentujące go dane są godne zaufania[6].

Czy musimy się przejmować mylącymi konotacjami pojęcia istotności statystycznej? Zdecydowanie tak. Co najmniej od czasów Stroopa [1935] wiadomo, że znaczenia słów podlegają przetwarzaniu automatycznemu i nie sposób ich ignorować. Konotacje

[6] Usłyszałem kiedyś, że osoby badane rzekomo popełniają mniej błędów w rozumowaniu, gdy jego przedmiotem nie są wielkości abstrakcyjne, lecz pieniądze. Z powodzeniem wykorzystałem tę informację w dydaktyce wnioskowania statystycznego. Przedstawiam na przykład istotę popularnej a nieprawidłowej interpretacji wartości *p* na przykładzie rozmowy dwóch osób, z których jedna, raczej pochopnie, godzi się podjąć pracę, dysponując jedynie informacją o tym, że jej oczekiwane zarobki – z około 95-procentową pewnością – nie będą zerowe, a druga w analogicznej sytuacji radośnie oczekuje „jeszcze większego" zysku, ponieważ wie, że jej zarobki nie wyniosą zero z aż 99-procentową pewnością. Niestety nie udało mi się zweryfikować pierwotnej informacji u źródła. Możliwe, że moja heurystyka dydaktyczna powstała wskutek nieporozumienia, odnoszę jednak wrażenie, że działa…

przymiotnika „istotny" są dekodowane i aktywowane w umyśle odbiorcy niezależnie od jego woli czy poziomu przygotowania statystycznego. Można się więc spodziewać, że skłonność do przydawania nieuzasadnionej emfazy wynikom opatrzonym etykietą istotności dotyczy, w mniejszym lub większym stopniu, *wszystkich* użytkowników testów istotności – nawet tych, którzy radzą sobie z zawiłościami prawdopodobieństw warunkowych. Ryzyko nadinterpretacji byłoby mniejsze, gdybyśmy trzymali się oryginalnej konwencji fisherowskiej i zamiast pisać na przykład: „skuteczność metody terapeutycznej *A* okazała się istotnie większa od skuteczności metody *B*", orzekali bardziej adekwatnie: „zaobserwowano przewagę skuteczności *A* nad *B* wystarczającą, by wstępnie odrzucić hipotezę o braku różnicy w populacji". Takie ostrożne sformułowanie daje bardziej rzetelny obraz faktycznej wagi pozytywnego wyniku testu istotności, zabezpieczając przed przecenianiem znaczenia opisywanej obserwacji nie tylko czytelnika, ale i samego badacza.

Znaczenie konotacji pojęcia istotności statystycznej uzmysławia prosty eksperyment: w części poświęconej dyskusji wyników pierwszego lepszego doniesienia z badań należy zastąpić wszystkie wystąpienia przymiotnika „istotny" frazą „prawdopodobnie niezerowy w populacji". Zamiast statystycznie istotnej zmiany przekonań uzyskamy zmianę prawdopodobnie niezerową; zamiast statystycznie istotnego wzrostu poziomu wykonania – wzrost prawdopodobnie większy niż zero; statystycznie istotna korelacja samopoczucia z temperamentem stanie się korelacją prawdopodobnie niezerową. Prosta sztuczka słowna radykalnie zmienia wydźwięk dyskusji. W przeciwieństwie do orzeczenia istotności statystycznej, które zwykle jest retorycznym zamknięciem, konstatacja niezerowej wartości efektu otwiera problem – zachęca do rzetelnej replikacji na większej próbie i prowokuje do stawiania notorycznie pomijanych pytań o wielkość i praktyczne znaczenie opisywanego efektu.

Wielkość efektu

Meehl [1978] zwracał uwagę, że badane efekty rzadko bywają idealnie zerowe, a zatem hipoteza zerowa jest w ścisłym sensie prawie zawsze fałszywa. Warto jednak zauważyć, że w praktyce psychologicznych badań ilościowych stosuje się zazwyczaj próby niewystarczające do wykrycia efektów małej, a często także średniej wielkości. Badacz sprawdza więc nie tyle, czy hipoteza zerowa jest fałszywa, tzn. czy efekt jest różny od zera, ale czy jest wystarczająco duży, by dało się go wykryć testem małej mocy. Test istotności staje się w takim użyciu niedoskonałym narzędziem do badania wielkości efektu. Znajduje to odzwierciedlenie w języku prac psychologicznych, rozróżniającym liczne odcienie istotności statystycznej. Autorzy piszą o efektach „zdecydowanie nieistotnych", „na granicy istotności", „wysoce istotnych" itp. Czasem odnoszą te określenia do pewności wniosku, czasem do wielkości efektu, a najczęściej do obu naraz: wartości p zdecydowanie większe od konwencjonalnego minimum $\alpha = 0{,}05$ traktują jako sygnał prawdopodobnej przypadkowości; wartości bliskie tego progu, ale go nie przekraczające – jako znak możliwości istnienia niezerowego efektu, którego jednak nie udało się zadowalająco potwierdzić; wartości równe lub niższe od progu uznaje się zaś za potwierdzenie efektu, którego wiarygodność i oczekiwana wielkość są tym

większe, im bardziej rygorystyczne kryterium α udało się spełnić. Mimo że opisany sposób interpretacji wartości *p* jest popularny, trudno go uznać za właściwy. Metodolodzy podkreślają często, że *p* nie jest miarą wielkości efektu, jednak ich głos nie znajduje właściwego odbicia w praktyce badawczej. Zapewne dlatego, że nie jest do końca prawdziwy: w typowych przypadkach, gdy liczebności prób i wariancje są zbliżone, a stosowane testy nie różnią się mocą, popularny „wskaźnik wielkości efektu" w postaci liczby zer po przecinku wartości *p* wykazuje całkiem rozsądną użyteczność praktyczną. Oczywiście z tego, że coś da się zrobić, wcale nie wynika, że należy. Przeciwnie: trudno rekomendować wnoszenie o istnieniu efektu w populacji, wielkości tego efektu i pewności wniosku na podstawie tylko jednej wartości – prawdopodobieństwa *p*, które zależy od arbitralnie przyjmowanych parametrów, jest (jak pokazują opisane wcześniej badania) błędnie rozumiane przez większość użytkowników, a do tego ze względu na swoją niestabilność [Cumming, 2014; Halsey i in., 2015] charakteryzuje się niewielką rzetelnością. Preferowanym rozwiązaniem jest ograniczenie stosowania testów istotności do ich podstawowej roli, czyli weryfikacji hipotez, i szacowanie wielkości efektów innymi metodami. Według zaleceń statystycznej *Task Force* Amerykańskiego Towarzystwa Psychologicznego testy istotności powinny być rutynowo uzupełniane informacją na temat wielkości opisywanych efektów, przy czym dla efektów kluczowych należy obliczyć także przedziały ufności [Wilkinson, APA Task Force on Statistical Inference, 1999].

Powściągliwa fisherowska interpretacja istotności statystycznej, w przypadku odrzucenia hipotezy zerowej, poprzestaje na konstatacji, że efekt w populacji jest prawdopodobnie różny od zera, co naturalnie prowokuje pytanie o wielkość efektu. Fałszywe rozumienie istotności przeciwnie – zniechęca do takiego pytania. Badacz, który nie odróżnia istotności statystycznej od praktycznej albo traktuje wynik istotny jak certyfikat zezwalający na proste przenoszenie wyników z próby na populację, przecenia wartość testu i nie widzi potrzeby oceny wielkości efektów. Sztafaż – egzotycznych dla humanisty – procedur statystycznych utrudnia mu dostrzeżenie, że opierając się na wynikach samego testu istotności, postępuje niewiele rozsądniej, niż gdyby podpisywał umowę kupna-sprzedaży, specyfikującą jedynie przybliżony poziom ufności co do tego, że zapłata będzie różna od zera.

Replikacje

Traktowanie pozytywnego wyniku testu istotności jako statystycznego certyfikatu realności badanego efektu jest – jak widzieliśmy – niewłaściwe. Tymczasem przykład, którego Fisher używa w swoim podręczniku, wydaje się przedstawiać właśnie taki przypadek. Statystyk opisuje „eksperyment psychofizyczny" sprawdzający zasadność twierdzenia pewnej damy, rzekomo zdolnej odróżnić smak herbaty z mlekiem sporządzanej tak, że do filiżanki z odrobiną mleka nalano herbaty, od takiej, w której do filiżanki z herbatą dodano mleka [Fisher, 1971, s. 11–12]. Zadaniem owej damy było posmakowanie herbaty z ośmiu filiżanek, z których połowę przygotowano według jednej, a połowę według drugiej recepty, i zdecydowanie, które cztery sporządzono w który sposób. Ponieważ istnieje 70 sposobów zestawienia grup po cztery z ośmiu

elementów, szanse prawidłowego zidentyfikowania wszystkich czterech filiżanek czystym przypadkiem wynoszą 1/70, czyli około 0,014. Szanse przypadkowego wyboru trzech właściwych i jednej niewłaściwej filiżanki są już znacząco większe i wynoszą 16/70, czyli około 0,23. O ile więc wynik 3 : 1 raczej nie wystarczyłby do odrzucenia hipotezy zerowej, o tyle wynik 4 : 0 jest już dość mało prawdopodobny, by uznać, że nie wziął się z przypadku. Choć sam Fisher o tym nie wspomina, wiadomo, że opisał rzeczywistą osobę i sytuację, a eksperyment został faktycznie przeprowadzony, z takim właśnie stuprocentowo poprawnym wynikiem [Salsburg, 2013].

Test – nazywany dziś testem dokładnym Fishera – wykazał istotną rozbieżność między wynikiem otrzymanym a przewidywanym przez hipotezę zerową. Wydawałoby się więc, że dostarczył tym samym potwierdzenia zasadności twierdzenia, zasłużonej dla rozwoju wnioskowania statystycznego, miłośniczki herbaty. Fisher pisze jednak tak:

> Badacze często przyjmują wartość 5 procent jako wygodny standard poziomu istotności, co oznacza gotowość ignorowania tych wyników, które owego standardu nie spełniają, a tym samym eliminowania z dalszej dyskusji większości fluktuacji wyników, wywołanych czynnikami losowymi. Nie da się takiej selekcji zrealizować w sposób, który wyeliminowałby wszystkie możliwe efekty przypadkowych koincydencji. Natomiast jeśli zaakceptujemy ową wygodną konwencję i uznamy zdarzenie, które czystym przypadkiem pojawia się tylko w jednej na 70 prób, za zdecydowanie „istotne" w sensie statystycznym, akceptujemy jednocześnie fakt, że *żadne oderwane doświadczenie, jakkolwiek by samo nie było istotne, nie stanowi wystarczającej doświadczalnej demonstracji zjawiska naturalnego*. Przysłowiowy „jeden przypadek na milion" zdarzy się bowiem niezawodnie, z nie większą i nie mniejszą od oczekiwanej częstością, jak wielkie by nie było nasze zdziwienie, gdy przytrafi się akurat nam. *By uznać jakieś zjawisko naturalne za doświadczalnie demonstrowalne, potrzebujemy nie izolowanej obserwacji, lecz sprawdzonej metody postępowania*. Co do testu istotności, można uznać zjawisko za demonstrowalne doświadczalnie, jeśli potrafimy przeprowadzić doświadczenie, które tylko rzadko nie da wyniku istotnego statystycznie[7].

Trudno o bardziej dobitne przedstawienie tej kwestii. Żaden pojedynczy wynik nie może być podstawą orzeczenia realności badanego efektu. Choćby najbardziej istot-

[7] „It is usual and convenient for experimenters to take 5 per cent, as a standard level of significance, in the sense that they are prepared to ignore all results which fail to reach this standard, and, by this means, to eliminate from further discussion the greater part of the fluctuations which chance causes have introduced into their experimental results. No such selection can eliminate the whole of the possible effects of chance coincidence, and if we accept this convenient convention, and agree that an event which would occur by chance only once in 70 trials is decidedly "significant," in the statistical sense, we thereby admit that no isolated experiment, however significant in itself, can suffice for the experimental demonstration of any natural phenomenon; for the "one chance in a million" will undoubtedly occur, with no less and no more than its appropriate frequency, however surprised we may be that it should occur to us. In order to assert that a natural phenomenon is experimentally demonstrable we need, not an isolated record, but a reliable method of procedure. In relation to the test of significance, we may say that a phenomenon is experimentally demonstrable when we know how to conduct an experiment which will rarely fail to give us a statistically significant result" [Fisher, 1971, s. 13–14, wyróżnienia P.W.].

na statystycznie, pojedyncza obserwacja jest tylko uzasadnieniem dalszej eksploracji. Mocniejsze wnioski wymagają replikacji [zob. też Wojciszke, 2004]. W ujęciu Fishera test istotności pełni funkcję daleko skromniejszą od tej, którą mu zwykle przydzielamy. Ma charakter bardziej przesiewowy niż konfirmacyjny. Jego zadanie można porównać do roli sita poszukiwacza złota, które mechanicznie odsiewa drobinki niewarte uwagi, ale nie zastępuje użytkownika w decydowaniu, która z pozostałych na siatce grudek jest, a która nie jest bryłką kruszcu.

BIBLIOGRAFIA

Bem, D.J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, *100*(3), 407–425. doi:10.1037/a0021524

Bem, D.J., Utts, J., Johnson, W.O. (2011). Must psychologists change the way they analyze their data? *Journal of Personality and Social Psychology*, *101*(4), 716–719. doi:10.1037/a0024777

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

Fenton, N., Neil, M. (2011). Avoiding probabilistic reasoning fallacies in legal practice using bayesian networks. *Australian Journal of Legal Philosophy*, *36*, 114.

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

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

Haller, H., Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research Online*, *7*(1).

Halsey, L.G., Curran-Everett, D., Vowler, S.L., Drummond, G.B. (2015). The fickle P value generates irreproducible results. *Nature Methods*, *12*(3), 179–185. doi:10.1038/nmeth.3288

Kalinowski, P., Fidler, F., Cumming, G. (2008). Overcoming the inverse probability fallacy. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *4*(4), 152–158. doi:10.1027/1614-2241.4.4.152

Kirk, R.E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, *56*(5), 746.

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.

Oakes, M.W. (1986). *Statistical Inference. A Commentary for the Social and Behavioural Sciences* (s. 185). New York: Wiley.

Salsburg, D. (2013). *The Lady Tasting Tea. How Statistics Revolutionized Science in the Twentieth Century*. New York: Henry Holt and Company.

Stroop, J.R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*, 643–662.

Wagenmakers, E.J., Wetzels, R., Borsboom, D., van der Maas, H.L. (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem (2011). *Journal of Personality and Social Psychology*, *100*(3), 426–432. doi:10.1037/a0022790

Westover, M.B., Westover, K.D., Bianchi, M.T. (2011). Significance testing as perverse probabilistic reasoning. *BMC Medicine*, *9*, 20. doi:10.1186/1741-7015-9-20

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.