Rocznik Kognitywistyczny 9/2016, s. 27–35
doi: 10.4467/20843895RK.16.003.5471
www.ejournals.eu/Rocznik-Kognitywistyczny

PIOTR WOLSKI

Istotność statystyczna I. Nieodrobiona lekcja

Statistical significance I

Abstract: In the present essay, the first in a short cycle, the author reviews and comments on the problems students and researchers have with proper understanding of the basics of statistical inference. Those difficulties seem to be in part due to mixing of the opposing theoretical stances of Fisher and Neyman, reviewed shortly. The author believes that the inconsistent standards of statistical inference afflict the teaching of methodology particularly.

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

W niniejszym, pierwszym z krótkiego cyklu esejów poświęconych praktyce stosowania testów istotności w badaniach psychologicznych autor omawia i komentuje dane demonstrujące trudności z prawidłowym rozumieniem podstaw wnioskowania statystycznego wśród studentów i badaczy. Jedną z przyczyn owych trudności wydaje się pomieszanie elementów, omówionych pokrótce, konkurencyjnych podejść Fishera oraz Neymana. Niespójne standardy inferencji statystycznej odbijają się niekorzystnie na procesie dydaktycznym.

Alarmująca diagnoza

Haller i Krauss [2002] przeprowadzili na wydziałach psychologii sześciu niemieckich uniwersytetów badania rozumienia podstaw wnioskowania statystycznego. Przedstawili respondentom przykład zastosowania jednego z najpopularniejszych narzędzi statystycznych, testu *t* Studenta i – podając sześć możliwych interpretacji uzyskanego wyniku ($t = 2.7$, $df = 18$, $p = 0{,}01$) – poprosili o zaznaczenie tych, które są w jego świetle logicznie uprawnione:

1. Dowiodłaś/-eś, że hipoteza zerowa jest bezwzględnie fałszywa (to jest między średnimi w populacji nie ma różnicy).
2. Dowiedziałeś/-aś się, jakie jest prawdopodobieństwo, że hipoteza zerowa jest prawdziwa.
3. Dowiodłaś/-eś, że hipoteza eksperymentalna (zakładająca istnienie różnicy między średnimi w populacji) jest bezwzględnie prawdziwa.

4. Możesz wywnioskować, jakie jest prawdopodobieństwo, że hipoteza eksperymentalna jest prawdziwa.
5. Jeśli postanowisz odrzucić hipotezę zerową, wiesz, jakie jest prawdopodobieństwo, że będzie to zła decyzja.
6. Twoje dane eksperymentalne są rzetelne w tym sensie, że gdyby – hipotetycznie – powtórzyć omawiany eksperyment bardzo wiele razy, uzyskałoby się statystycznie istotny wynik w 99% przypadków.

Wszystkie wymienione stwierdzenia są fałszywe. Najbardziej oczywiste jest to w przypadku zdań 1 i 3. Test istotności opiera się na danych z próby i ma z natury charakter probabilistyczny, nie może więc być podstawą żadnych definitywnych rozstrzygnięć w kwestii parametrów populacji. Zdania 2 i 4 sugerują, że test pozwala oszacować prawdopodobieństwo hipotezy zerowej $P(H_0)$ lub jego przeciwieństwo, $P(H_1) = 1 - P(H_0)$. W procedurze testowania istotności nie szacuje się jednak prawdopodobieństwa hipotezy zerowej, tylko bada konsekwencje założenia jej prawdziwości, a to nie to samo. Badacz oblicza, jakie byłoby, w opisywanym przez hipotezę zerową wypadku braku efektu w populacji, prawdopodobieństwo pojawienia się w próbie efektu równego faktycznie zaobserwowanemu lub większego[1]. Jeśli owo prawdopodobieństwo (p) jest wystarczająco niskie ($p \leq \alpha$), badacz uznaje model populacji określony hipotezą alternatywną za lepiej pasujący do danych i odrzuca hipotezę zerową. W przeciwnym wypadku powstrzymuje się od decyzji[2]. Wniosek o populacji ma więc, jak widać, charakter jakościowy, a nie ilościowy: Na podstawie uzyskanych danych badacz podejmuje decyzję, czy jego obserwacja jest wystarczająco wiarygodna, by uznać, że nie wzięła się z przypadku. Choć wartość p informuje go o stopniu niezgodności danych z hipotezą zerową, badacz nie wie dokładnie, jakie są szanse, że podjął właściwą decyzję, ani ryzyko, że zdecydował źle.

Zdanie 5 mówi o – ocenianym po odrzuceniu hipotezy zerowej – ryzyku, że decyzja okaże się błędna. Ponieważ odrzucenie hipotezy zerowej jest błędem, gdy hipoteza ta jest prawdziwa, omawiane ryzyko wynosi $P(H_0)$ i, jak wspominaliśmy, nie jest znane. Należy odrzucić także zdanie 6, bowiem prawdopodobieństwa replikacji nie sposób ocenić bez wiedzy, czy efekt istnieje w populacji, czy nie. W pierwszym przypadku oczekiwana w długiej serii frakcja istotnych wyników równa się nieznanej mocy testu ($1 - \beta$), w drugim – prawdopodobieństwu błędu pierwszego rodzaju, α.

Haller i Krauss [2002] wzorowali swój sondaż na wcześniejszym projekcie Oakesa [1986], który badał rozumienie podstaw wnioskowania statystycznego wśród psychologów akademickich. Badacze niemieccy ułatwili swoim respondentom zadanie, podpowiadając w instrukcji, że prawdziwych może być kilka, ale też i żadne z podanych stwierdzeń. Poszerzyli także grupę badanych psychologów o studentów i nauczycieli metodologii.

[1] Testy istotności pozwalają badać zgodność zaobserwowanych danych z dowolną hipotezą zerową, także przyjmującą niezerową wartość parametru populacji. W praktyce badań psychologicznych testuje się jednak przeważnie hipotezy, zakładające zerową wartość efektu eksperymentalnego – wyrażonego jako różnica między statystykami (średnimi, proporcjami, liczebnościami) albo wartość statystyki, obrazującej siłę związku (współczynnik korelacji liniowej, wartość β w analizie regresji). W trosce o komunikatywność częściowo rezygnuję ze ścisłości i czasem opisuję testy istotności tak, jakby ograniczały się tylko do tej dominującej klasy.

[2] Taką, asymetryczną regułę decyzyjną przyjmuje klasyczny model Fishera.

Wyniki Oakesa były zaskakująco złe: niemal wszyscy badani (68 z 70 osób) popełnili przynajmniej jeden z sześciu możliwych błędów interpretacyjnych. Mimo ułatwienia w postaci podpowiedzi w instrukcji replikacja Hallera i Kraussa (N = 113) wykazała podobnie alarmujący poziom niezrozumienia podstaw testów istotności – przynajmniej jeden błąd popełniło 100% ankietowanych studentów, 90% badaczy i 80% nauczycieli statystyki! Autorzy pocieszają się, że trzej uczestniczący w badaniu profesorowie odpowiedzieli bezbłędnie…

Rozdałem kilka razy studentom pierwszego roku studiów doktoranckich na psychologii ankiety zawierające zestaw zdań Oakesa – z analogicznym skutkiem. Mimo instrukcji podpowiadającej, że żadne ze zdań może nie być prawdziwe, nikt z ankietowanych nie wybrał takiej możliwości. Przyznaję to niezbyt chętnie, bo większość z tych studentów sam starałem się – pięć lat wcześniej – nauczyć m.in. właściwej interpretacji testów istotności. Ankietę wypełniali uczestnicy dobrowolnego repetytorium, którzy siłą rzeczy nie czuli się pewnie w tej problematyce. Z pewnością na kursie zaawansowanym byłoby lepiej, jednak nie sądzę, by nasi absolwenci stanowili wyjątek od reguły powszechnej trudności z prawidłowym rozumieniem logiki testów istotności.

We wszystkich grupach badanych przez Hallera i Kraussa zdecydowanie najczęściej uznawano za prawdziwe twierdzenie nr 5, dotyczące prawdopodobieństwa fałszywego alarmu. Akceptowało je aż siedmiu na dziesięciu nauczycieli metodologii i badaczy. Podobnie było w badaniu Oakesa oraz w moim własnym sondażu.

Prawdopodobieństwo, o którym mowa w zdaniu nr 5, może się wydawać znane, bowiem przyjmowany w testach poziom istotności to właśnie prawdopodobieństwo fałszywego alarmu. W pośpiechu badania ankietowego łatwo ulec heurystyce poznawczej, bazującej na poczuciu znajomości, i nie dostrzec, że poziom istotności to prawdopodobieństwo odrzucenia hipotezy zerowej, *gdy* jest ona prawdziwa, podczas gdy w ocenianym zdaniu mowa o prawdopodobieństwie, *że* odrzucona hipoteza zerowa jest prawdziwa, a to nie to samo. Trudno jednak podobnie tłumaczyć akceptację zdań 2 i 4, postulujących znajomość $P(H_0)$ lub $P(H_1)$. Jeśli się prawidłowo rozumie logikę weryfikacji hipotez statystycznych, to nawet w pośpiechu trudno uznać, że test istotności dostarcza tych informacji. Zdanie 2 zostało tymczasem uznane za prawdziwe przez 17% nauczycieli metodologii, 26% badaczy i 32% studentów, a zdanie 4 przez 33% metodologów oraz badaczy i 59% studentów. Akceptacja tych zdań w badaniach Oakesa była jeszcze wyższa: odpowiednio 36% i 66% ankietowanych badaczy.

Ze statystyką na bakier

Miałem parę razy okazję opowiadać psychologom – badaczom i studentom – o wynikach Oakesa oraz Hallera i Kraussa. Najczęściej reagowali uśmiechem lekkiego zażenowania, jakby to była zasłużona reprymenda za nieodrobioną lekcję. Sam przy czytaniu pracy Hallera i Kraussa odczuwałem rodzaj wstydu i poczucia zbiorowej winy.

Czy te reakcje są zasadne? Po części zapewne tak. Psychologia przyciąga głównie osoby, których zainteresowania i zdolności leżą w domenie szeroko rozumianej humanistyki. Co zrozumiałe, typowy student naszej dyscypliny przykłada się bardziej do zdobywania wiedzy z zakresu psychologii społecznej, dziecka czy klinicznej niż do

zgłębiania podstaw wnioskowania statystycznego. Problemy metodologiczne wydają mu się też zwykle – zwłaszcza na początku studiów – odległe i abstrakcyjne. Niewystarczająca motywacja czy brak predyspozycji nie mogą być jednak jedyną, ani nawet główną, przyczyną powszechnego niezrozumienia podstaw weryfikacji hipotez statystycznych, ujawnionego przez badania Hallera i Kraussa. Niemal tak samo często jak studenci błędnych odpowiedzi udzielali bowiem także pracownicy i nauczyciele statystyki, a ci przecież nie stronią od abstrakcyjnego myślenia. Badacze są też silnie zmotywowani do nauki warsztatu, bez znajomości którego nie mogą efektywnie prowadzić badań ani publikować wyników. Może więc problem tkwi nie tylko w nieodrobionej lekcji, ale też w jej przedmiocie?

Kirk, autor podręczników, na których wychowało się kilka pokoleń psychologów, pisze:

> Testowanie hipotez to skomplikowany rytuał, trudny do zrozumienia i nauczania. Co semestr widzę, jak zdziwienie maluje się na twarzach początkujących studentów, którym tłumaczę, że badacze testują hipotezy zerowe, które uważają za fałszywe, z nadzieją że zostaną one odrzucone, co uprawdopodobni hipotezy alternatywne, które ich zdaniem są prawdziwe [Kirk 2001, s. 215].

Faktycznie, studentom nie jest łatwo zrozumieć, dlaczego chcąc sprawdzić, czy dane potwierdzają istnienie telepatii, powinni pytać, czy to prawda, że nieprawda, że telepatia nie istnieje. Już pobieżny kontakt z wynikami prac naukowych uczy ich, że wynik jest istotny statystycznie, czyli relatywnie wiarygodny, jeśli wartość p – cokolwiek by ona znaczyła – jest mniejsza od 0,05, ale próby pogłębienia i usystematyzowania wiedzy na temat testowania istotności są zwykle frustrujące. Według podręcznika poziomem istotności nazywa się prawdopodobieństwo fałszywego alarmu α, ale w tabelach nagłówek „istotność" opisuje często kolumnę z wartościami p. Czy więc poziom istotności to p, czy α? Za wysoce istotne uznaje się te wyniki, dla których p oraz α są małe. Czy to ma znaczyć, że istotność jest tym większa, im jej poziom jest niższy? To zagadkowy oksymoron. Dlaczego wynik o niewielkim obciążeniu ryzykiem błędu nazywa się istotnym lub znaczącym (ang. *significant*), skoro on w istocie jest tylko wiarygodny? Przecież znaczenie i wiarygodność to niezależne wymiary: nikt nie uznaje bagatelnego faktu za istotny tylko dlatego, że go ustalił ponad wszelką wątpliwość. A jednak badacze używają często pojęcia istotności w sposób sugerujący, że nie chodzi im o wiarygodność, ale właśnie o wagę i znaczenie opisywanych wyników. Pytania się mnożą: czemu w publikacjach czasem stosuje się zapis ze znakiem nierówności $p \leq 0{,}05$; czasem ze znakiem równości, np. $p = 0{,}02$; ale pewnych zapisów, takich jak np. $p \leq 0{,}02$ nie spotyka się wcale (choć już $p \leq 0{,}01$ pojawia się często)? Dlaczego autor podręcznika sugeruje, że jeśli $p \leq \alpha$, to należy przyjąć hipotezę alternatywną (H_1), ale w przeciwnym wypadku, zamiast przyjąć hipotezę zerową (H_0), każe powstrzymać się od decyzji, skoro wcześniej istotę weryfikacji hipotez przedstawiał inaczej? Miał to być wybór pomiędzy alternatywami H_0 i H_1, w którym prawdopodobieństwo niesłusznego odrzucenia H_0 i przyjęcia H_1 to błąd pierwszego rodzaju α, a prawdopodobieństwo niesłusznego odrzucenia H_1 i przyjęcia H_0 to błąd drugiego rodzaju, β. Czy więc test istotności jest tylko próbą falsyfikacji hipotezy zerowej, czy wyborem między hipotezą zerową a alternatywną? To ważne pytania i nie tylko studenci miewają kło-

poty z uzyskaniem jasnych odpowiedzi. Zatrzymajmy się przy ostatnim: falsyfikacja jednej hipotezy czy konfirmacja jednej z dwóch alternatyw? Popularne podręczniki zwykle nie informują, że pierwsze ujęcie wywodzi się od Ronalda Fishera, twórcy najpopularniejszej w psychologii odmiany metody weryfikacji hipotez statystycznych, a drugie to zmodyfikowany wariant, opracowany przez głównego adwersarza Fishera, Jerzego Neymana, we współpracy z Egonem Pearsonem.

Sir Ronald Aylmer Fisher, statystyk i doświadczalnik, uważał, że czas badacza jest zbyt cenny, by ten marnował go na zajmowanie się efektami o wątpliwej rzetelności, których iluzoryczność ujawniają, dopiero poniewczasie, nieudane próby replikacji. Opracował więc procedurę pozwalającą na częściowe odsianie takich losowych fluktuacji [Fisher 1971]. Posiłkując się logiką podobną do zastosowanej przez Karla Pearsona[3] w teście dobroci dopasowania, sprawdzał, czy wielkość zaobserwowanego w próbie efektu odbiega znacząco od wartości oczekiwanej w sytuacji, w której ów efekt byłby dziełem przypadku. Przyjął, że tymczasowe założenie nieobecności efektu w populacji, które nazwał „hipotezą zerową", można odrzucić, jeśli zaobserwowane dane są z nim znacząco sprzeczne, tzn. jeśli przy braku efektu w populacji prawdopodobieństwo *p* wystąpienia w próbie efektu co najmniej takiego jak zaobserwowany byłoby odpowiednio niskie. W niniejszym eseju skupiam się na – najpopularniejszych w badaniach psychologicznych – testach istotności różnic oraz związku. W odniesieniu do nich odrzucenie hipotezy zerowej oznacza robocze przyjęcie, że odpowiednik testowanego efektu ma w populacji wartość różną od zera[4]. Istotę podejścia Fishera podsumowuje często cytowane zdanie z jego wpływowej monografii: „O każdym doświadczeniu można powiedzieć, że istnieje tylko po to, by dać faktom szansę zaprzeczenia hipotezie zerowej" [Fisher 1971, s. 16].

Schemat fisherowski zakłada asymetryczną regułę decyzyjną. Jeśli wynik doświadczenia jest znacząco sprzeczny z hipotezą zerową, można przyjąć, że efekt różni się od zera nie tylko w próbie, ale i w populacji. Jednak w przeciwnym wypadku – np. gdy *p* nie wynosi 0,01, a 0,1 – nie można wyniku uznać za potwierdzenie hipotezy zerowej, bowiem jest on z nią tylko nie dość sprzeczny, a to nie to samo. W myśl popularnej formuły retorycznej brak dowodu nie jest dowodem braku.

[3] Za pierwszy przypadek zastosowania testu istotności uważa się argument Johna Arbuthnota, który w 1712 roku użył prostego wariantu testu znaków do odrzucenia hipotezy, zakładającej, że obserwowana corocznie, w danych o liczbie chrzcin chłopców i dziewcząt, przewaga tych pierwszych wynika z przypadkowych odchyłek od, uważanej przez niego błędnie za „losową", proporcji 1 : 1. Wyliczone, znikomo małe, prawdopodobieństwo tej hipotezy uznał za powód do jej odrzucenia, a tym samym mocny argument na rzecz teorii boskiego planu, w którym wyższa śmiertelność chłopców miałaby być zawczasu kompensowana większą liczbą ich urodzin [Hald 2003]. Pierwszym szerzej stosowanym testem istotności był jednak dopiero test chi-kwadrat Karla Pearsona [Gigerenzer 1989].

[4] Nawet wtedy, gdy hipoteza zerowa nie zakłada zerowej wartości parametru populacji – np. w teście zgodności frakcji zaobserwowanej z przewidywaną, można powiedzieć, że zakłada ona zerową wartość efektu: Jeśli pretendent do miana jasnowidza systematycznie odgaduje wynik rzutu kostką z częstością $\varphi = 1/6$, uznamy to za przejaw zerowych zdolności, bowiem ich faktyczną miarą (E) nie jest owa częstość obserwowana, lecz różnica (zerowa) między nią a częstością oczekiwaną – nawet jeśli model formalny zakłada hipotezę H_0: $\varphi = 1/6$, a nie H_0: $E = \varphi - 1/6 = 0$.

Jerzy Neyman był innego zdania [Neyman, Pearson 1928a; 1928b]. Uważał, że procedura weryfikacji hipotez powinna umożliwiać symetryczne, jednoznaczne rozstrzygnięcie między hipotezą przypadku H_0 a jej alternatywą H_1. Jeśli założone graniczne kryterium istotności wynosi α, dla $p \leq \alpha$ badacz powinien odrzucić H_0 i przyjąć H_1, a dla $p > \alpha$ przeciwnie: przyjąć H_0 i odrzucić H_1. By taka reguła decyzyjna była zasadna, akceptowalnie niskie musi być jednak nie tylko ryzyko błędu pierwszego rodzaju (fałszywego alarmu) α, minimalizowane w procedurze Fishera, ale także nie mniej ważne ryzyko błędu drugiego rodzaju (przeoczenia), β. Neyman pozwala więc badaczowi dowodzić braku, ale wymaga, by jego procedura miała wystarczającą zdolność wykrywania istniejących efektów, tzw. moc statystyczną, 1 – β. W praktyce oznacza to m.in. konieczność prowadzenia badań na odpowiednio dużych próbach[5]. Wariant Neymana różni się od wersji Fishera podejściem do prawdopodobieństwa *p*. Dla Fishera *p* jest źródłem informacji o sile, z jaką dane przemawiają przeciw hipotezie zerowej. Mimo istnienia popularnej konwencji α = 0,05 badacz może i powinien bardziej wierzyć wynikowi, dla którego $p = 0{,}003$, niż temu, dla którego poziom zaobserwowanej istotności wynosi $p = 0{,}048$. Nie powinien też traktować tej ostatniej sytuacji jako znacząco różnej od przypadku $p = 0{,}053$. Według Neymana zaś badacz musi przyjąć i konsekwentnie utrzymywać w kolejnych badaniach stałe kryterium α. Tylko wtedy bowiem prawidłowa jest interpretacja α jako oczekiwanej przy prawdziwej H_0 częstości fałszywych alarmów w długiej serii powtórzeń.

Fisher uważał, że Neyman nie rozumie natury nauk empirycznych, w których nie ma miejsca dla mechanicznego, czarno-białego rozstrzygania o rzeczywistości. Dostrzegał przydatność jego metody do algorytmizowania niektórych procesów decyzyjnych – np. w przemysłowej kontroli jakości, gdzie ryzyko błędu wnioskowania jest założone z góry, a stosunek zysku z trafnej decyzji do kosztów błędu policzalny i łatwy do optymalizacji [Fisher 1971] – jednak pomysł stosowania jej w postępowaniu naukowym uważał za całkowicie chybiony. Znany z trudnego charakteru matematyk krytykował swojego kolegę bez pardonu przy każdej nadarzającej się okazji. Neyman zastanawiał się, dlaczego raz, w czasie konferencji we Francji, Fisher odstąpił od utartego schematu i wysłuchawszy referatu, nie wypowiedział ani jednej krytycznej uwagi. Wyjaśnienie okazało się prozaiczne – jego adwersarz nie znał francuskiego [Salsburg 2013].

Zasługi Neymana i Pearsona są daleko większe, niż był to skłonny przyznać Fisher. Najważniejszą z nich jest zwrócenie uwagi na problem mocy statystycznej, czyli zdolności testu do wykrywania istniejących efektów [Neyman, Pearson 1933]. Niestety, owa moc jest w badaniach psychologicznych z reguły zbyt mała [Button i in. 2013; Cohen 1962; Sedlmeier, Gigerenzer 1989; Vankov, Bowers, Munafò 2014], by dało się zastosować model symetrycznego wyboru między hipotezą zerową a alternatywną. W takich przypadkach nie ma alternatywy dla interpretacji testów istotności według reguły Fishera.

W praktyce badawczej pomieszanie elementów podejść Fishera i Neymana prowadzi do rozmaitych niekonsekwencji. Jasno widać je na przykładzie testowania

[5] Fisher miał świadomość, że – podobnie jak doskonalenie planów eksperymentalnych – zwiększanie liczebności prób pozwala na uzyskiwanie danych o większym ładunku informacyjnym, jednak sformalizowanie pojęcia mocy statystycznej jest zasługą Neymana.

założenia normalności rozkładu. Mimo iż większość stosowanych testów istotności odwołuje się do logiki fisherowskiej, w myśl której interpretowanie wyniku nieistotnego jako dowodu nieobecności efektu w populacji jest błędem, testy normalności opierają się na schemacie bliższym założeniom Neymana-Pearsona: hipotezę zerową (założenie normalności rozkładu cechy w populacji) przyjmuje się, jeśli $p > 0{,}05$. Niestety, przy tym najczęściej nie sprawdza się mocy statystycznej, która w przypadku zmiennych o niewielkiej liczbie obserwacji, tak częstych w badaniach psychologicznych, bywa bardzo niska. Ta niekonsekwencja oczywiście podkopuje wiarygodność całego testu. Nie dowierzamy przecież wnioskowi kogoś, kto szukał i nie znalazł, jeśli wiemy, że szukał byle jak[6].

Prace Neymana i Pearsona przyczyniły się do ogromnego spopularyzowania testów istotności i uczyniły z nich w wielu dziedzinach – w tym psychologii – złoty standard. Mniej znany jest fakt, że sam Neyman nie był entuzjastą tego obrotu spraw. Świadom ograniczeń weryfikacji hipotez, z czasem coraz chętniej sięgał po metody estymacyjne [Salsburg 2013]. Mimo to paradoksalnie pamiętamy Jerzego Neymana bardziej jako reformatora fisherowskiej procedury weryfikacji hipotez niż matematyka, który wymyślił przedziały ufności [Zieliński 2009].

Widoczny nie tylko w wynikach Oakesa oraz Hallera i Kraussa, alarmująco niski poziom rozumienia podstaw wnioskowania statystycznego ma wiele przyczyn. Metodolodzy wskazują na co najmniej trzy grupy: długo upowszechniane niewłaściwe standardy [Brzeziński 2012; Finch i in. 2004; Finch, Thomason, Cumming 2002; Wilkinson, APA Task Force on Statistical Inference, 1999], niedostatki sposobu nauczania metod wnioskowania statystycznego [Gliner, Leech i Morgan 2002; Haller i Krauss 2002], oraz braki samych metod [Cumming 2014; Gigerenzer 1989; 2004; Ioannidis 2005; Meehl 1978; Westover, Westover i Bianchi 2011].

O tym, jak dalece niezadowalający jest zdaniem krytyków stan rzeczy w naszej dyscyplinie, dają pojęcie już same tytuły niektórych znanych prac: *Theoretical Risks and Tabular Asterisks: Sir Karl, Sir Ronald, and the Slow Progress of Soft Psychology* [Meehl 1978]; *The earth is round (*$p < .05$*)* [Cohen 1994]; *Significance Tests Die Hard: The Amazing Persistence of a Probabilistic Misconception* [Falk i Greenbaum 1995]; *Mindless statistics* [Gigerenzer 2004]; *Why most published research findings are false* [Ioannidis 2005]; *Significance testing as perverse probabilistic reasoning* [Westover i in. 2011][7]. Gigerenzer pisał ze swadą:

> Przyszłym historykom psychologii będzie trudno zrozumieć osobliwy rytuał, udający *sine qua non* metody naukowej, który pojawił się w latach 50. i był praktykowany do samego końca dwudziestego wieku. W podręcznikach psychologicznych i pedagogicznych owego okresu znajdą oni rozmaite nazwy

[6] Twórca testu dobroci dopasowania, chi-kwadrat, Karl Pearson uważał jego interpretowanie w kategoriach weryfikacji hipotez o populacji za całkowicie nieuprawnione. Uważał też, że jego wyniki są znaczące tylko dla dużych prób o wielkości rzędu 100 [Inman 1994].

[7] „Ryzyko teoretyczne gwiazdek w tabelach: Sir Karl, Sir Ronald i powolny postęp miękkiej psychologii"; „Ziemia jest okrągła ($p < 0{,}05$)"; „Testy istotności umierają powoli: Zadziwiająca trwałość probabilistycznego nieporozumienia"; „Statystyka bez głowy"; „Czemu większość publikowanych wyników badań jest fałszywa"; „Testowanie istotności jako perwersja rozumowania probabilistycznego".

> tego rytuału: „wnioskowanie statystyczne", „testowanie hipotez zerowych", „testowanie istotności", czy ostatnio „NHSTP" [*Null-Hypothesis Significance-Test Procedure*]. Ze zdziwieniem dowiedzą się, że rytuał został szybko zinstytucjonalizowany, mimo iż: (1) wpływowi ówcześni psychologowie – w tym Sir Frederick Bartlett, R. Duncan Luce, Herbert Simon, B.F. Skinner oraz S.S. Stevens zgodnie sprzeciwiali się jego stosowaniu [...]; (2) statystycy Sir Ronald Fisher, Jerzy Neyman oraz Egon S. Pearson odrzuciliby NHSTP jako niekonsekwentny zlepek swoich idei [...]; (3) mało który ze znanych statystyków tamtych czasów był jej zwolennikiem; i (4) choć w psychologii uznano ją za warunek naukowości, owa metoda nigdy nie przyjęła się w naukach przyrodniczych[8] [Gigerenzer, 1998, s. 199].

Dalej autor porównał procedurę testowania istotności do natręctwa, zmuszającego pacjentów do ciągłego mycia rąk, dodając, że wielu autorów podręczników oraz większość stosujących tę procedurę badaczy w istocie nie rozumie jej głównego wyniku – wartości *p*. To naprawdę bardzo mocne słowa.

Haller i Krauss [2002] piszą, że kłopoty z właściwym rozumieniem podstaw wnioskowania statystycznego to problem, który nauczyciele dzielą z uczniami. Ci ostatni są jednak w dużo gorszej sytuacji: Ostatnią rzeczą, która przychodzi do głowy – skołowanemu licznymi sprzecznościami i niejasnościami – studentowi czy studentce, jest podejrzenie, że praktyka stosowania metod wnioskowania statystycznego w psychologii grzeszy niekonsekwencją, terminologii brak ścisłości, założenia najpopularniejszej procedury wnioskowania statystycznego opierają się na błędzie logicznym, a jej wyniki są powszechnie źle rozumiane nawet przez nauczycieli. Raczej dochodzi ona czy on do przedwczesnego wniosku, że statystyka jest za trudna i rezygnuje z prób jej głębszego zrozumienia. Przyswaja kilka uproszczonych reguł, resztę kuje na pamięć, a po egzaminie szybko zapomina, zostając z poczuciem byle jak odrobionej lekcji.

BIBLIOGRAFIA

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

Button, K.S., Ioannidis, J.P., Mokrysz, C., Nosek, B., Flint, J., Robinson, E.S., Munafò, M.R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Review Neuroscience*, *14*(5), 365–376. doi:10.1038/nrn3475

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, *65*, 145–53.

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

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

[8] To uogólnienie idzie nieco za daleko, bo choć faktycznie w fizyce, astronomii czy geografii rzadko używa się testów istotności, to już w biologii jest inaczej: Fisher opracowywał metody analizy statystycznej głównie z myślą o swojej wczesnej pasji, rolnictwie.

Falk, R., Greenbaum, W. (1995). Significance tests die hard: The amazing persistence of a probabilistic misconception. *Theory & Psychology*, *5*(1), 75–98. doi:10.1177/0959354395051004

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.

Finch, S., Thomason, N., Cumming, G. (2002). Past and future American Psychological Association guidelines for statistical practice. *Theory and Psychology*, *12*, 825–853.

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

Gigerenzer, G. (1989). *The Empire of Chance: How Probability Changed Science and Everyday Life*. Cambridge–New York: Cambridge University Press.

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

Gliner, J.A., Leech, N.L., Morgan, G.A. (2002). Problems with null hypothesis significance testing (NHST): What do the textbooks say? *The Journal of Experimental Education*, *71*(1), 83–92.

Hald, A. (2003). *A History of Probability and Statistics and Their Applications before 1750*. Hoboken, NJ: John Wiley & Sons, Inc.

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

Inman, H.F. (1994). Karl Pearson and RA Fisher on statistical tests: A 1935 exchange from Nature. *The American Statistician*, *48*(1), 2–11.

Ioannidis, J.P. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), 696–701. doi:10.1371/journal.pmed.0020124

Kirk, R.E. (2001). Promoting good statistical practices: Some suggestions. *Educational and Psychological Measurement*, *61*(2), 213–218.

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.

Neyman, J., Pearson, E.S. (1928a). On the use and interpretation of certain test criteria for purposes of statistical inference: Part II. *Biometrika*, 263–294.

Neyman, J., Pearson, E.S. (1928b). On the use and interpretation of certain test criteria for purposes of statistical inference: Part I. *Biometrika*, 175–240.

Neyman, J., Pearson, E.S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *231*(694–706), 289–337. doi:10.1098/rsta.1933.0009

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

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

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

Vankov, I., Bowers, J., Munafò, M.R. (2014). On the persistence of low power in psychological science. *Quarterly Journal of Experimental Psychology (Hove)*, *67*(5), 1037–1040. doi:10.1080/17470218.2014.885986

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.

Zieliński, R. (2009). Przedział ufności dla frakcji. *Matematyka Stosowana*, *10*, 1–17.