ACTA UNIVERSITATIS LODZIENSIS
FOLIA PSYCHOLOGICA 19, 2015

http://dx.doi.org/10.18778/1427-969X.19.01

Wiesław Szymczak

Uniwersytet Łódzki
Wydział Nauk o Wychowaniu, Instytut Psychologii
Zakład Metodologii Badań Psychologicznych i Statystyki
91-433 Łódź, ul. Smugowa nr 10/12
e-mail: wieszym@uni.lodz.pl

POJĘCIE WIELKOŚCI EFEKTU NA TLE TEORII NEYMANA–PEARSONA TESTOWANIA HIPOTEZ STATYSTYCZNYCH[1]

Abstrakt. Celem tej pracy jest zwrócenie uwagi badaczy wykorzystujących metody statystyczne w analizie wyników swoich badań na pomieszanie dwóch różnych teorii testowania hipotez statystycznych, teorii Fishera i teorii Neymana–Pearsona. Zawarcie, w obecnie stosowanym instrumentarium statystycznym, pomysłów z obu tych teorii, powoduje, że znakomita większość badaczy bez chwili namysłu za prawdziwą przyjmuje stwierdzenie, iż im mniejsze prawdopodobieństwo, tym silniejsza zależność. Przedstawione zostały słabe strony teorii Neymana–Pearsona i wynikające z nich problemy przy podejmowaniu decyzji w wyniku przeprowadzonych testów. Problemy te stały się usprawiedliwionym poszukiwaniem mniej zawodnych rozwiązań, jednakże zaproponowane mierniki wielkości efektu, jako wykorzystujące z jednej strony dogmat o związku między wielkością prawdopodobieństwa w teście i siłą zależności, a z drugiej – brak jakichkolwiek podstaw teoretycznych tego rozwiązania, wydają się jeszcze jednym pseudorozwiązaniem rzeczywiście występujących problemów. Dodatkowo, wykorzystywanie mierników wielkości efektów wygląda na próbę zwolnienia badaczy z głębokiego myślenia o uzyskanych wynikach z analizy statystycznej, w kategoriach merytorycznych. Powstał trywialny przepis: odpowiednia wartość miernika natychmiast implikuje siłę zależności – podejście takie wydaje się niegodne badacza.

Słowa kluczowe: teorie testowania hipotez statystycznych, prawdopodobieństwo, moc testu, empiryczna moc testu, wielkość efektu.

1. WPROWADZENIE

Wśród badaczy stosujących metody statystyczne (a dokładniej: testujących hipotezy statystyczne i podejmujących decyzje na podstawie rezultatów testów) do opracowywania wyników swoich badań stosunkowo często można spotkać następującą opinię: im mniejsze prawdopodobieństwo w teście, tym istotniejszy

[1] Artykuł ten składa się z fragmentów przygotowywanej do druku książki na temat wnioskowania statystycznego.

wynik (silniejsza zależność). Na ile prawdziwe jest to stwierdzenie i z czego ono wynika? Otóż jest ono konsekwencją pomieszania dwóch różnych podejść do teorii testowania hipotez statystycznych, teorii Fishera i teorii Neymana–Pearsona.

Problem badaczy wykorzystujących w opracowaniu wyników badań ilościowych metody testowania hipotez, niekiedy nawet nieuświadomiony, polega na tym, że w praktyce wszystkie stosowane testy statystyczne są tzw. testami istotności, tj. testami, które nie kontrolują prawdopodobieństwa błędu drugiego rodzaju. Wszystkie one kontrolują prawdopodobieństwo błędu pierwszego rodzaju, lecz nie kontrolując prawdopodobieństwa błędu drugiego rodzaju, uniemożliwiają podjęcie decyzji o przyjęciu hipotezy zerowej. Jeśli prawdopodobieństwo w teście jest większe od przyjętego poziomu istotności (najczęściej jest to wartość $\alpha = 0{,}05$), stwierdzamy, że nie ma podstaw do odrzucenia hipotezy zerowej. Praktycznie jesteśmy wówczas w sytuacji pełnej niewiedzy. Nieco lepiej, choć też nie w sposób doskonały, wygląda sytuacja, gdy prawdopodobieństwo w teście jest mniejsze od przyjmowanego poziomu istotności. Podejmujemy wówczas decyzję o odrzuceniu hipotezy zerowej (traktujemy ją jako fałszywą) i przyjęciu hipotezy alternatywnej (uznajemy ją za prawdziwą).

Ale i w tym przypadku również nie mamy komfortowej sytuacji. Uznajemy, że relacje czy zależności opisane hipotezą alternatywną są prawdziwe, jednak badacza zazwyczaj zaczyna wówczas interesować, jak silne są to relacje.

Dość powszechna interpretacja, że im mniejsze prawdopodobieństwo uzyskane w teście, tym silniejsza zależność (tutaj w terminach merytorycznych), nie ma żadnego uzasadnienia statystycznego. Badaczy ciągle dręczy pytanie: „jak silna jest ta zależność?". Pytanie to można potraktować jako szczególną wersję ogólniejszego problemu: czy wnioskowanie statystyczne (*statistical inference*) i wnioskowanie naukowe (*scientific inference*) są tym samym. Zagadnienie to ciągle jeszcze nie zostało rozwiązane i jest przyczyną dyskusji między statystykami i badaczami stosującymi statystykę.

W dalszej części artykułu spróbuję wyjaśnić powody obecnych problemów z interpretacją wyników testowania hipotez statystycznych oraz rzeczywiste niedoskonałości istniejących rozwiązań. Informacje te pozwolą Czytelnikowi uświadomić sobie, dlaczego pojawiło się coś takiego, jak pojęcie wielkości efektu oraz czym skutkuje jego wykorzystywanie.

Analiza statystyczna nie zajmuje się badaniem zjawisk deterministycznych, jej przedmiotem są zjawiska losowe. Aby w pewien sposób „okiełznać" nieprzewidywalność pojawiania się takich zdarzeń, niezbędna jest pewna miara, pozwalająca – z lepszym lub gorszym skutkiem – przewidywać nieprzewidywalne. Taką miarą w statystyce, przynajmniej na pierwszym etapie jej rozwoju, było prawdopodobieństwo. Kłopot z tą miarą polega na tym, że nie posiadamy intuicji prawdopodobieństwa. Skutkuje to np. takimi stwierdzeniami: „Jeśli prawdopodobieństwo jakiegoś zdarzenia jest prawie równe 1, to z dużym stopniem pewności zdarzenie to pojawi się w pojedynczej próbie" (Papoulis, 1972). Papoulis

pokazuje tym stwierdzeniem, na czym polega problem z prawdopodobieństwem. Bo cóż oznacza duży stopień pewności? Jest to po prostu inna nazwa prawdopodobieństwa. Zatem cytowane zdanie nic nie wyjaśnia. I musimy się zgodzić, że „teoria statystyczna, która jest ścisłą dyscypliną rozwiniętą z jasno sformułowanych aksjomatów, jest powiązana ze zjawiskami fizycznymi tylko poprzez nieścisłe terminy" (Papoulis, 1972). Jednakże trudno zgodzić się z opinią, że statystyka jest dyscypliną rozwiniętą z jasno sformułowanych aksjomatów. Raczej różne statystyki są rozwijane z jasno sformułowanych różnych zbiorów aksjomatów.

Ale wróćmy do zagadnień prawdopodobieństwa zdarzenia. Brak intuicji prawdopodobieństwa zdarzenia spowodował powstanie wielu definicji prawdopodobieństwa, co doskonale utrudnia późniejsze ich wykorzystanie w analizach statystycznych.

2. PRAWDOPODOBIEŃSTWO

Poniżej przedstawię cztery definicje prawdopodobieństwa: definicję aksjomatyczną, definicję wykorzystującą częstości względne (von Mises), definicję klasyczną i prawdopodobieństwo jako miara przekonania.

Najbardziej nośną i najefektywniejszą okazała się aksjomatyczna definicja prawdopodobieństwa sformułowana przez Kołmogorowa w 1933 r. Jest ona do dzisiaj podstawą wszelkich rozważań probabilistycznych.

2.1. Definicja aksjomatyczna (Kołmogorow, 1933)

Każdemu zdarzeniu (zdarzeniu losowemu) A przyporządkowana jest liczba $P(A)$, spełniająca następujące warunki:

1) $P(A)$ jest nieujemna; $P(A) \geq 0$,
2) prawdopodobieństwo zdarzenia pewnego jest równe jedności; $P(\Omega) = 1$,
3) prawdopodobieństwo alternatywy (sumy mnogościowej) skończonej lub przeliczalnej ilości zdarzeń losowych parami wyłączających się jest równe sumie prawdopodobieństw tych zdarzeń:

$$P\left(\bigcup_k A_k\right) = \sum_k P(A_k); \quad A_i \cap A_j = \varnothing \quad i, j = 1, 2, \ldots, k; \quad i \neq j \tag{1}$$

Wzór ten można zapisać w nieco innej postaci:

$$P(A_1 \cup A_2 \cup \cdots \cup A_k \cup \cdots) = P(A_1) + P(A_2) + \ldots + P(A_k) + \ldots$$

$$A_i \cap A_j = \varnothing \quad i, j = 1, \ldots, k; \quad i \neq j \tag{2}$$

Oprócz własności prawdopodobieństwa wynikających bezpośrednio z aksjomatycznej definicji, czyli własności, iż prawdopodobieństwo zdarzenia pewnego jest równe jedności:

$$P(\Omega) = 1 \tag{3}$$

oraz że prawdopodobieństwo alternatywy (sumy mnogościowej) skończonej lub przeliczalnej ilości zdarzeń losowych parami wyłączających się jest równa sumie prawdopodobieństw tych zdarzeń, warto dodać jeszcze jedną: prawdopodobieństwo zdarzenia niemożliwego jest równe zero:

$$P(\varnothing) = 0 \tag{4}$$

Tak zdefiniowane prawdopodobieństwo w żaden sposób nie poprawia intuicji tego pojęcia. Jest wygodne, eleganckie i efektywne dla rozwijanej na jego podstawie teorii probabilistycznej, lecz nie ułatwia (a nawet nie umożliwia) interpretacji podczas oceny rezultatów analiz statystycznych.

2.2. Klasyczna definicja prawdopodobieństwa (Laplace, 1812)

Klasyczna definicja prawdopodobieństwa sformułowana przez Laplace'a znajduje zastosowanie tylko w przypadku skończonych zbiorów zdarzeń elementarnych.

Jeśli przestrzeń zdarzeń elementarnych Ω składa się z n zdarzeń elementarnych (wyników doświadczenia losowego) **jednakowo możliwych** i jeżeli wśród nich jest *k* zdarzeń elementarnych sprzyjających zajściu zdarzenia A, to liczbę:

$$P(A) = \frac{k}{n}, \tag{5}$$

nazywamy prawdopodobieństwem zajścia zdarzenia *A*. Prawdopodobieństwo zdarzenia *A*, zgodnie z tą definicją, znajdujemy a priori bez przeprowadzania doświadczenia.

Pewnego wyjaśnienia może wymagać zwrot „zdarzenia elementarne sprzyjające zajściu zdarzenia *A*". Rozważmy zdarzenie polegające na wyrzuceniu nieparzystej liczby oczek, w eksperymencie polegającym na rzucie sześcienną kostką do gry. W tej sytuacji wyrzucenie ścianki z jednym oczkiem albo z trzema albo z pięcioma oczkami będzie powodowało, iż uznamy, że zaszło interesujące nas zdarzenie (nieparzysta liczba oczek). Zatem każde ze zdarzeń elementarnych {•; •••; •••••} będzie zdarzeniem sprzyjającym zajściu zdarzenia *A*.

Klasyczna definicja prawdopodobieństwa ma dwie poważne wady. Pierwsza to założenie, że wszystkie zdarzenia elementarne muszą być jednakowo

możliwe, inaczej mówiąc – muszą być jednakowo prawdopodobne, zatem w definicji prawdopodobieństwa używamy już pojęcia prawdopodobieństwa. Drugi problem to wymaganie, by przestrzeń zdarzeń elementarnych składała się ze skończonej liczby elementów. Gdy zbiór Ω jest nieskończony to n nie jest liczbą skończoną i iloraz k/n nie daje się obliczyć nawet wtedy, gdy k jest liczbą skończoną. Wówczas zamiast liczby elementów musimy używać innych liczb, zwanych miarami zbiorów, pełniących podobną rolę jak liczebności, lecz będzie to już inna definicja.

2.3. Definicja wykorzystująca częstości względne (von Mises, 1936)

Rozpatrywane doświadczenie przeprowadzane jest wielokrotnie, np. n razy. Wśród n wyników doświadczenia zdarzenie A pojawiło się n_A razy (n_A razy pojawiło się zdarzenie elementarne sprzyjające zajściu zdarzenia A). Doświadczenie to wykonujemy dalej. Teoretycznie można sobie wyobrazić, że nieskończoną ilość razy. Wówczas prawdopodobieństwo zdarzenia A można interpretować jako:

$$P(A) = \lim_{n\to\infty} \frac{n_A}{n} \tag{6}$$

Oznacza to, że jeśli eksperyment losowy (doświadczenie losowe) będziemy wykonywać wielokrotnie i po każdym wykonaniu eksperymentu obliczać częstość badanego zdarzenia A, to wraz ze wzrostem liczby wykonanych doświadczeń wahania częstości n_A/n będą coraz mniejsze i będą oscylować wokół pewnej stałej wartości, wokół liczby będącej prawdopodobieństwem $P(A)$. Lecz, niestety, nie możemy utożsamiać częstości – nawet obliczonej na podstawie ogromnej liczby przeprowadzonych doświadczeń – z prawdopodobieństwem zdarzenia. Dlatego też często tę definicję prawdopodobieństwa traktuje się jako tzw. częstościową interpretację prawdopodobieństwa, bardzo wygodną do celów stosowania statystyki matematycznej. Interpretacja ta znajduje również zastosowanie, gdy przestrzeń zdarzeń elementarnych zawiera nieskończoną ilość elementów.

2.4. Prawdopodobieństwo jako miara przekonania (prawdopodobieństwo subiektywne)

Prawdopodobieństwo tego rodzaju używane jest jako miara przekonania, że coś może albo nie może być prawdą; jak prawdopodobne jest konkretne zdarzenie. Oczywiście jest to subiektywna ocena orzekającego o wielkości prawdopodobieństwa i nie jest ono oparte na jakichkolwiek obliczeniach. Jednakże, jako prawdopodobieństwo, jest nie mniejsze od zera i nie większe od jedności.

Wydawać by się mogło, że ze względu na swój subiektywizm pojęcie tego prawdopodobieństwa nie znajdzie zastosowania. Nic bardziej błędnego – jest ono przedmiotem wielu artykułów, nie tylko z zakresu zastosowań, lecz także teorii, np. Anscombe i Aumann (1963), w którym znalazł się rozdział o istnieniu subiektywnych prawdopodobieństw, Machina i Schmeidler (1992) czy Karni (1993).

3. KONKURENCYJNE TEORIE TESTOWANIA HIPOTEZ STATYSTYCZNYCH

Aby zrozumieć istotę kontrowersji wokół testowania hipotez, musimy zapoznać się z dwiema konkurencyjnymi teoriami: teorią Fishera i teorią Neymana–Pearsona. Postępowanie według teorii Fishera nazywane bywa wnioskowaniem indukcyjnym (*inductive inference*) zaś według teorii Neymana–Pearsona postępowaniem indukcyjnym (*inductive behavior*). Obie te teorie zostały zaproponowane w latach 30. XX w. (Fisher, 1935; Neyman, Pearson, 1933) i wprowadzają one całkowicie różne metodologie. Zagadnienia wielu, nie tylko Fishera i Neymana–Pearsona, „teorii statystyki" były, i nadal są, przedmiotem zainteresowania matematyków i statystyków (Inman, 1994; Lehmann, 1993, 1995; Berger, 2003; Christensen, 2005; Manthey, 2010).

3.1. Teoria Fishera

W podejściu Fishera formułowana jest tylko jedna hipoteza – hipoteza zerowa, H_0, która odpowiada skonstruowanemu modelowi badawczemu. Testowanie tej hipotezy polega na wybraniu pewnej statystyki testowej T o znanym rozkładzie prawdopodobieństwa oraz obliczeniu jej wartości na podstawie wyników próby. Duża wartość statystyki T, a więc małe prawdopodobieństwo p odpowiadające tej wartości, dostarczała badaczowi dowodów przeciwko H_0. Dostatecznie mała wartość p powodowała odrzucenie hipotezy H_0. Fisher swoją procedurę testowania uzasadniał tym, że wartość p (*p-value*) może być traktowana jako „siła dowodu" przeciwko H_0 (*"strength of evidence" against* H_0). Mała wartość p wskazywała mało prawdopodobne zdarzenie, a w konsekwencji czyniła mało prawdopodobnym prawdziwość hipotezy badanej i doprowadzała do jej odrzucenia.

3.2. Teoria Neymana–Pearsona

Neyman i Pearson oprócz hipotezy zerowej zaproponowali hipotezę alternatywną. **Zarówno hipoteza zerowa, jak i alternatywna były hipotezami prostymi**, np.:

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta = \theta_1 \end{cases} \tag{7}$$

Sposób postępowania autorów podczas testowania hipotezy był następujący:

– odrzucenie H_0, jeśli $T \geq c$ i zaakceptowanie alternatywnej H_1; przyjęcie H_0, gdy $T < c$, gdzie c jest z góry ustaloną wartością krytyczną testu,

– obliczenie prawdopodobieństw błędów pierwszego i drugiego rodzaju, $\alpha = P_0$(odrzucenia H_0) i $\beta = P_1$(zaakceptowania H_0).

Uzasadnieniem Neymana dla tej procedury była częstościowa interpretacja prawdopodobieństwa, czyli, w wielokrotnie powtarzanych badaniach z użyciem tej samej procedury statystycznej, częstość podjęcia błędnej decyzji polegającej na odrzuceniu prawdziwej hipotezy zerowej nie powinna być większa, niż określone z góry prawdopodobieństwo (Neyman, 1977). Neyman i Pearson całkowicie rozwiązali problem testowania w przypadku prostej hipotezy zerowej i prostej hipotezy alternatywnej (lemat Neymana–Pearsona). Jednak dla bardziej złożonych przypadków testowania, np. złożonych hipotez alternatywnych, teoria wymagała dodatkowych pomysłów. Opracowywanie szczegółów rozwiązywania złożonych zagadnień testowania było głównym przedmiotem zainteresowań statystyki matematycznej (teoretycznej) w następnych dekadach.

3.3. Nieco szczegółów wynikających z teorii Neymana–Pearsona

Dlaczego teoria Neymana–Pearsona? Otóż podejście Neymana–Pearsona, nazywane też podejściem częstościowym, a niekiedy nawet ortodoksyjnym (Dienes, 2011), mimo różnych „zanieczyszczeń" przeniesionych z teorii Fishera oraz krytyki wielu użytkowników, pozostaje najczęściej wykorzystywaną metodą testowania hipotez statystycznych.

Gwoli przypomnienia, jaką hipotezę nazywamy prostą, a jaką złożoną – określenia te formułowane są w różny sposób. I tak:

– hipotezę nazywamy prostą, gdy określa ona jednoznacznie rozkład prawdopodobieństwa; każda hipoteza, która nie jest prostą nazywa się złożoną (Neyman, 1969),

– hipoteza statystyczna jest prosta, czyli pojedyncza, albo złożona stosownie do tego, czy zawiera jeden punkt czy wiele punktów (także punkt w przestrzeni wielowymiarowej) (Zubrzycki, 1970),

– hipoteza H, precyzująca wartość wszystkich nieznanych parametrów, nosi nazwę hipotezy prostej. Hipoteza niespełniająca tego warunku nosi nazwę hipotezy złożonej (Fisz, 1969).

Oczywiście zarówno hipoteza zerowa, jak i hipoteza alternatywna może być prosta lub złożona, lecz w praktyce nie jest to już takie oczywiste.

W podejściu częstościowym wykorzystujemy częstościową interpretację prawdopodobieństwa: przy wielokrotnym powtarzaniu procedury statystycznej i podejmowaniu wynikających z niej decyzji, częstość błędnych decyzji nie będzie większa niż przyjęte z góry prawdopodobieństwo. Ostatnie stwierdzenie w praktyce odnosi się tylko do prawdopodobieństwa α.

A dlaczego nie do β? W sformułowaniu Neymana–Pearsona, w problemie testowania występują dwie hipotezy proste. Natomiast praktyka testowania hipotez statystycznych jest zupełnie inna. Mamy, co prawda, do czynienia z zerową hipotezą, która jest hipotezą prostą, ale hipoteza alternatywna jest prawie zawsze złożona.

Tu natychmiast pojawia się pytanie: dlaczego to hipoteza zerowa ma być prosta, a alternatywna złożona? Nie musi tak być. Na przykład Rao (1982) rozważa sytuacje, w których zarówno hipoteza zerowa, jak i alternatywna są hipotezami złożonymi:

H_0	H_1
$\Theta \leq \theta_0$	$\Theta > \theta_0$
$\Theta \leq \theta_0$ lub* $\theta \geq \theta_1$	$\Theta_0 < \theta < \theta_1$
$\Theta_0 \leq \theta \leq \theta_1$	$\Theta < \theta_0$ lub* $\theta > \theta_1$

*Zamiast „lub" poprawniejszy jest w tym przypadku łącznik „albo".

Problem z tak sformułowanymi hipotezami polega na wyznaczeniu takiej funkcji φ ($\alpha(\varphi) = E(\varphi|\theta)$), żeby wielkość $\alpha(\theta')$ osiągała maksimum dla $\theta' \in H_1$ przy warunku:

$$\alpha(\theta) \leq \alpha \quad \text{dla} \quad \theta \in H_0 \tag{8}$$

W ogólnym przypadku zadanie to może nie mieć zadowalającego rozwiązania, tym samym nie uzyskamy właściwego testu.

Silvey (1978) przedstawia „matematyczną" metodę przezwyciężenia trudności wynikających z ewentualnych nieciągłości rozkładów prawdopodobieństwa. Omawiana przez niego metoda w ogólnym przypadku, ale dotyczącym testowania prostej hipotezy zerowej przeciwko prostej alternatywnej, prowadzi do testu najmocniejszego na poziomie istotności α. Jednak w sytuacji prostej hipotezy zerowej i złożonej alternatywnej test jednostajnie najmocniejszy nie istnieje. Zilustrowane jest to następującym przykładem.

Niech $x_1, x_2, \ldots, x_n$ będzie próbką losową z rozkładu normalnego o wariancji równej 1. Na podstawie takich obserwacji testujemy zagadnienie:

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta \neq \theta_0 \end{cases} \tag{9}$$

Dla takiego zagadnienia nie istnieje test jednostajnie najmocniejszy. Jak poradzić sobie w takiej i podobnych sytuacjach? „Moglibyśmy spróbować na drodze rozważań heurystycznych znaleźć jakąś ogólną metodę konstrukcji testów i rozwiązać dany problem tą właśnie metodą, licząc przy tym na to, że chociaż być może uzyskane rozwiązanie nie znajdzie uzasadnienia w świetle dotychczasowych

kryteriów, to jednak doprowadzi do testu, który w sposób właściwy, choć niekoniecznie optymalny, wykorzystuje informacje zawarte w wynikach naszych obserwacji" (Silvey, 1978).

W większości praktycznych zastosowań testów statystycznych używamy „intuicyjnie sensownych testów". Takim testem jest powszechnie znany test *t*-Studenta porównywania dwóch wartości oczekiwanych, przedstawiony w przykładzie 1.

Przykład 1. Zagadnienie porównywania dwóch wartości oczekiwanych dla prób niezależnych.

$$\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 \neq \mu_2 \end{cases} \tag{10}$$

Niech $(x_{11}, x_{12}, \ldots, x_1 n_1)$, $(x_{21}, x_{22}, \ldots, x_{2n2})$ będą wynikami pomiarów pewnej cechy X w próbach pobranych z dwóch rozłącznych populacji. Jeśli badana cecha ma rozkład normalny w każdej z tych dwóch podpopulacji oraz wariancje tejże cechy są jednakowe w tych podpopulacjach (choć nie znamy ich wartości), a ponadto prawdziwa jest hipoteza zerowa, to statystyka:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{n_1 \cdot s_1^2 + n_2 \cdot s_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \tag{11}$$

ma rozkład *t*-Studenta z $n_1 + n_2 - 2$ stopniami swobody (Zubrzycki, 1970). Dwustronny test *t*-Studenta jest testem jednostajnie najmocniejszym nieobciążonym (Magiera, 2007).

Pomińmy pewne szczegóły z powyższych sformułowań (są one zrozumiałe jedynie przez statystyków teoretyków) i zastanówmy się, kiedy moglibyśmy skorzystać z testu *t*-Studenta w celu porównania dwóch wartości oczekiwanych. U podstaw powyższych twierdzeń (gdyż są to twierdzenia w sensie matematycznym, które zostały udowodnione; ale tylko w terminach teoretycznych) leżą trzy założenia. Powtórzę je:

– normalność rozkładu badanej cechy w każdej z dwóch podpopulacji,
– jednorodność wariancji badanej cechy w każdej z dwóch podpopulacji,
– prawdziwość hipotezy zerowej o równości wartości oczekiwanych cechy X w obu podpopulacjach.

Jeśli te trzy założenia są spełnione, to wówczas statystyka (11) ma rozkład *t*-Studenta i test, który ją wykorzystuje, ma odpowiednie cechy dla testowania hipotez (10).

Blalock w swoim podręczniku (1975) stwierdza: „[…] stawiana (zerowa) hipoteza jest zwykle tą, którą chcemy odrzucić. […] W rzeczywistości spodziewamy się zwykle, że hipoteza zerowa jest błędna i mamy nadzieję odrzucić ją na korzyść hipotezy alternatywnej". Czyli, z praktycznego punktu widzenia, zależy

nam na tym, aby jedno z trzech założeń powyższych twierdzeń nie było spełnione. Ale będzie to skutkowało nieprawdziwością tezy. Podobne wnioski będą wynikały z niespełniania dwóch pozostałych założeń.

Cóż zatem będzie oznaczało używanie „intuicyjnie sensownych testów"? Jest to określenie zdecydowanie zbyt liberalne. Na czyjej to intuicji mamy polegać? Pewniej byłoby polegać na wiedzy i to na wiedzy dobrze ugruntowanej. Będzie nas to zmuszało do stosowania rozwiązań przybliżonych (tu pojawiają się problemy miary bliskości), asymptotycznych (a tu z kolei, problemy szybkości zbieżności), lecz znajdujących uzasadnienie w teorii.

Wróćmy jednak do zagadnienia mocy testu. Przez moc testu będziemy rozumieć zdolność testu (w terminach prawdopodobieństwa) do wykrycia fałszywości hipotezy zerowej w sytuacji, gdy jest ona rzeczywiście fałszywa. Nawet w sytuacji gdy hipoteza zerowa jest prosta i prostą jest hipoteza alternatywna, mamy do czynienia z dwupunktowym zbiorem parametrów wyznaczających dwa rozkłady prawdopodobieństwa. W przypadku złożonej hipotezy alternatywnej często zbiór parametrów ma moc continuum (moc zbioru liczb rzeczywistych). Zamiast więc mówić o mocy testu używa się pojęcia funkcji mocy. Funkcja mocy (*power function*) $\pi(\theta)$ określa prawdopodobieństwo podjęcia akcji odrzucenia H_0, które to prawdopodobieństwo jest funkcją parametru θ. Oprócz pojęcia funkcji mocy testu $\pi(\theta)$ używane jest pojęcie funkcji operacyjno-charakterystycznej (*operating characteristic*).

Mając do czynienia z prostą hipotezą zerową i prostą alternatywną, prawdopodobieństwo błędu pierwszego rodzaju i moc testu możemy opisać za pomocą funkcji mocy następująco:

$$\begin{aligned} \alpha &= \pi(H_0) = \Pr(\text{odrzucenie } H_0 \mid H_0 \text{ jest prawdziwa}) \\ 1-\beta &= \pi(H_1) = \Pr(\text{odrzucenie } H_0 \mid H_0 \text{ jestfaszywa}) \end{aligned} \tag{12}$$

Natomiast w sytuacji hipotez złożonych:

$$\begin{aligned} \alpha &= \max_{\theta \text{ w } H_0} P_0(\text{odrzucić } H_0) = \max_{\theta \text{ w } H_0} \pi(\theta) \\ \beta &= \max_{\theta \text{ w } H_1} [1-\pi(\theta)] \end{aligned} \tag{13}$$

Jeśli maksimum nie istnieje, to symbol max zastępujemy symbolem supremum (sup); α czasami nazywana jest poziomem istotności testu (*significance level of the test*) (Lindgren, 1962).

Formułując zagadnienie testowania według przesłanek Neymana-Pearsona, określamy hipotezę zerową (hipotezę prostą) i hipotezę alternatywną (która prawie zawsze jest hipotezą złożoną). Na przykład:

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases} \tag{14}$$

Najogólniej mówiąc, własności testu, w tym także jego moc, będą zależały od prawdziwej wartości parametru w H_1, a tej nie znamy. Konsekwencją tego problemu jest konstruowanie testów statystycznych kontrolujących prawdopodobieństwo błędu pierwszego rodzaju, a nie kontrolujących prawdopodobieństwa błędu drugiego rodzaju. Czytelnika zainteresowanego szczegółami funkcji mocy testu odsyłam do podręcznika Lindgrena (1962). Nie jest to może lektura najprostsza, ale rzetelna.

4. EMPIRYCZNA (OBSERWOWANA) MOC TESTU

Na początku należy postawić sobie pytanie, czy coś takiego jak empiryczna moc testu w ogóle istnieje. W świetle powyższych rozważań wydaje się, że nie istnieje. Cóż zatem jest obliczane w programach statystycznych? Na to pytanie jest bardzo trudno sensownie odpowiedzieć. Jak widzieliśmy w rozważaniach teoretycznych, moc testu zależy od wartości parametru zaszytego w hipotezie alternatywnej, której to wartości nie znamy. W praktyce uzależnia się moc testu od statystyki będącej podstawą testu, wielkości próby, wariancji zmiennej w populacji generalnej, wielkości różnicy między hipotezą zerową i prawdziwą hipotezą alternatywną, poziomu istotności testu i kierunkowości tego testu (Williams, Zimmerman, 1989). Jak widać z wyliczenia składowych mocy testu, najczęściej znana jest nam jedynie wielkość próby. O'Keefe (2007) ujmuje zagadnienie podobnie, choć je nieco upraszczając: cztery zmienne (moc, poziom istotności testu, wielkość próby i wielkość efektu w populacji) są związane w ten sposób, że gdy wartości trzech spośród nich są ustalone, to czwarta jest w pełni określona. I natychmiast autor stawia pytanie: zakładając, że badacz nie zna wielkości efektu w populacji generalnej, to jak może obliczyć moc testu? Odpowiada, iż moc jest liczona dla potencjalnej wielkości efektu w populacji generalnej. Zatem mówienie o mocy konkretnego testu statystycznego jest mylące. Jeszcze bardziej mylące jest mówienie o mocy testu *post hoc*. Moc testu jest taka sama bez względu na to, kiedy moc jest obliczana, przed czy po wykonaniu testu.

Hoenig i Heisey (2001) jeszcze raz zauważają od dawna znaną prawidłowość, iż istnieje ścisły związek między poziomem istotności testu i jego mocą. W przypadku prawdopodobieństwa uzyskanego w teście, większego od przyjętego poziomu istotności (badaną zależność uznamy za „nieistotną" ze statystycznego punktu widzenia), oszacowana moc testu będzie mała.

Jednak w piśmiennictwie pojawia się inna ciekawa zależność, której warto poświęcić kilka zdań. Mianowicie, czy istnieje jakiś związek między brakiem istotności uzyskanym w teście statystycznym a jakością badania? Nazywam to jakością badania, ale w literaturze angielskojęzycznej występuje pojęcie *power of study*, które można tłumaczyć jako „zdolność badania". Nie do końca odpowiada to jakości badania, ale dosłowne tłumaczenie jako „moc badania" brzmi niezgrabnie.

Po bliższym przyjrzeniu się powyższemu sformułowaniu wydaje się, że jest to po prostu przejęzyczenie, mała precyzja wypowiedzi. Nie udało mi się znaleźć definicji mocy badania. Obiecujący tytuł artykułu Sedlmeiera i Gigerenzera (1989): *Do Studies of Statistical Power Have an Effect on the Power of Studies* jest nieprecyzyjny, gdyż w artykule mówi się o *power studies*, czyli badaniach mocy, a nie o *power of studies*, czyli mocy badań. Zatem intrygujące zdanie o zależności mocy testu i mocy badania w rzeczywistości dotyczy jedynie poziomu istności i mocy tego samego testu statystycznego.

W wielu artykułach zamieszczane są tabele zawierające retrospektywne moce testu dla różnych zagadnień statystycznych, a więc dla różnych statystyk (np. Onwuegbuzie, Leech, 2004; Lenth, 2007). Ich przydatność, w świetle powyższych rozważań, wydaje się mocno wątpliwa.

A do czego jest nam potrzebne oszacowanie mocy testu statystycznego? Jak pamiętamy, prawie wszystkie stosowane w praktyce testy statystyczne są tzw. testami istotności, czyli testami niekontrolującymi prawdopodobieństwa błędu drugiego rodzaju. W takiej sytuacji uzyskując w teście prawdopodobieństwo większe od poziomu istotności, nie mamy podstaw do odrzucenia hipotezy zerowej i praktycznie jesteśmy w stanie pełnej niewiedzy, jaką możemy i powinniśmy podjąć decyzję. Znajomość mocy testu mogłaby ułatwić podjęcie odpowiedniej decyzji, np. przy dużej mocy testu moglibyśmy pokusić się o przyjęcie hipotezy zerowej. Lecz w praktyce to się raczej nie zdarzy, czyli ocena mocy testu *post hoc* wydaje się nieprzydatna.

5. KONTROWERSJE WOKÓŁ TESTOWANIA HIPOTEZY ZEROWEJ

Traktowanie wartości prawdopodobieństwa p jako miary dowodu przeciwko H_0 spowodowało powstanie poglądu, że im mniejsza wartość p, tym większa istotność dowodu (ale przeciwko hipotezie zerowej, a nie za hipotezą alternatywną, gdyż takiej w rozumowaniu Fishera nie ma). Po odrzuceniu hipotezy zerowej, i w konsekwencji odrzuceniu zaproponowanego modelu, badacz musi skonstruować inny model. Fisher często przekonywał, że jest ważne móc testować hipotezę zerową, nawet wtedy, gdy żadna hipoteza alternatywna nie została określona. Sensowność takiego postępowania była szeroko dyskutowana i wielu statystyków zdecydowanie ją popiera.

Żadna z tych teorii nie jest idealna, na każdej z nich ciążą poważne zarzuty. Teorii Neymana–Pearsona zarzuca się brak wrażliwości na zmienność siły dowodu przy odrzuceniu hipotezy zerowej. Hipoteza zerowa zostaje odrzucona zarówno dla, np. $t = 2$, jak i $t = 81$ przy $\alpha = 0{,}05$. Podejście Neymana–Pearsona krytykowane było również z powodu potrzeby określania hipotezy alternatywnej i w konsekwencji trudności z określeniem prawdopodobieństwa błędu drugiego.

Z kolei p w teorii Fishera było podstawą zarzutu naruszenia częstościowej zasady prawdopodobieństwa. Warto w tym miejscu przypomnieć, że praca

Kołmogorowa, w której przedstawił układ aksjomatów prawdopodobieństwa zdarzenia, ukazała się dopiero w roku 1933, a więc wydaje się, że w momencie powstawania teorii testowania hipotez statystycznych nie była jeszcze powszechnie znana. Jeffreys uważał, że logika wykorzystująca wartość *p* pod ogonem funkcji gęstości (w przeciwieństwie do rzeczywistych danych) jest głupia – „[…] hipoteza, która być może jest prawdziwa, może być odrzucona ponieważ nie przewidziano obserwowalnych rezultatów, które nie pojawiły się" (Jeffreys, 1961). W podobnym duchu wypowiadał się Fisz (1969). Nazywając testy stosowane w zagadnieniach testowania hipotezy zerowej (bez hipotezy alternatywnej) testami istotności, zauważa: „czy można uważać za udowodnione, że hipoteza H_0 jest niesłuszna, gdy prawdopodobieństwo zdarzenia [sformułowanego w H_0] jest bardzo małe? Otóż nie można, gdyż chociaż prawdopodobieństwo tego zdarzenia jest – przy słuszności hipotezy H_0 – bardzo małe, to jednak zdarzenie to może nastąpić" (W teorii miary mamy do czynienia ze zbiorami miary zero, a prawdopodobieństwo jest unormowaną miarą zbioru).

Teorie, z jednej strony Fishera, z drugiej Neymana i Pearsona, są całkiem różne. Znajduje to odzwierciedlenie w fakcie, że są dla nich używane odrębne określenia (mimo że czasami niekonsekwentnie): testowania istotności dla Fishera i testowania hipotez dla Neymana i Pearsona. (Ponieważ oba dotyczą testowania hipotez, więc często ignorowane są terminologiczne różnice i stosowany jest termin „testowanie hipotez" niezależnie od tego, czy testowanie jest przeprowadzane „na sposób" Fishera czy Neymana–Pearsona) (Lehman, 1993).

Szeroka gama poglądów dotyczących metodologicznych problemów wynikających z różnych teorii testowania hipotez statystycznych wyrażanych przez matematyków, statystyków, filozofów znajduje się w komentarzach do artykułu Bergera (2003). Dyskutanci podnoszą też zagadnienia unifikacji podejść, ale także utrzymania istniejącej różnorodności, przytaczając argumenty zarówno za, jak i przeciw unifikacji. Jednakże w większości opracowań współczesnej elementarnej statystyki elementy tych dwóch niekompatybilnych podejść są mieszane, co bardzo niekorzystnie odbija się na stosowaniu metod statystycznych w praktyce.

W świetle istnienia dwóch, niejako konkurencyjnych, teorii testowania hipotez statystycznych, naturalne wydaje się pytanie, czy rozumiemy co robimy, testując hipotezy statystyczne. To pytanie stawiają sobie już nie matematycy i statystycy, ale badacze stosujący metody statystyczne do opracowywania wyników badań ilościowych: psycholodzy, socjolodzy, specjaliści od zarządzania (Hubbard, Armstrong, 2006; Levine i in., 2008; Roberts, Pashler, 2000; Rodgers, 2010; Sterne, 2002; Denis, 2003; Jones, Tukey, 2000; Killeen, 2005; Thompson, 1994).

Thompson (1994) zauważa, że „zbyt mało badaczy rozumie, co testy statystyczne »robią a czego nie robią«" i w konsekwencji wyniki ich badań są błędnie interpretowane. Nawet jeśli badacz rozumie elementy testowania hipotez statystycznych, to nie jest to zintegrowane z jego badaniem. Na przykład, wpływ wielkości próbki na istotność statystyczną może zostać zauważony przez badacza, ale

to spostrzeżenie nie zostaje przekazane podczas interpretacji wyników badania, w którym mieliśmy wiele tysięcy elementów. Co prawda, problem tak licznej próby najczęściej nie dotyczy psychologów, ale dobrze jest zdawać sobie z niego sprawę. Choć nawet w psychologii mamy niekiedy do czynienia z bardzo dużymi próbami, np. w psychologii społecznej.

I jeszcze jeden cytat z Thompsona (1994): „Jako naukowcy, musimy zadawać pytania: a) jakie są efekty wynikające z wielkości próbki?, b) czy te rezultaty można uogólnić? Testowanie hipotez statystycznych nie udziela odpowiedzi na te pytania. Tak więc, testowanie hipotez statystycznych może odwracać uwagę od znacznie ważniejszych rozważań".

Teoria testowania hipotez Neymana–Pearsona z prawdopodobieństwem błędu pierwszego rodzaju α jako poziomem istotności testu jest powszechnie uznawana jako norma w metodologii testowania hipotez statystycznych. Jednak model Fishera testowania istotności, gdzie wyraźnie wartość p oznacza poziom istotności (ale nie jest to poziom istotności testu, tylko poziom istotności przeciwko prawdziwości hipotezy zerowej) zdominował praktykę testowania (Hubbard, Bayarri, 2003). Paradoks ten powstał z powodu rozbieżności (niezgodności) tych dwóch teorii, które w obecnie istniejącym podejściu do testowania zostały anonimowo – nikt się do tego nie przyznaje – wymieszane razem, tworząc fałszywe wrażenie jednego, spójnego modelu wnioskowania statystycznego (Hubbard, Armstrong, 2006).

Z powodów, które powyżej naszkicowałem, angielskojęzyczne zwroty *significance testing*, *statistical significance* w zasadzie nie zawierają żadnej treści. W języku polskim też spotykamy „testowanie istotności", „istotność statystyczną" czy „w sposób istotny statystycznie…", które, moim zdaniem, nie powinny być stosowane. Co prawda, sam nie jestem bez grzechu, gdyż zwrotu „w sposób istotny statystycznie…" jednak używam. Może te rozważania zwrócą uwagę badaczy na konieczność stosowania jednoznacznej, precyzyjniejszej terminologii statystycznej.

W podręczniku (Szymczak, 2010) używam pojęcia „testy istotności" dla testów statystycznych, w których nie kontrolujemy prawdopodobieństwa błędu drugiego rodzaju, polegającego na przyjęciu fałszywej hipotezy zerowej. Z powodu nieznajomości hipotezy alternatywnej (w znakomitej większości praktycznych zagadnień hipoteza alternatywna jest hipotezą złożoną) nie jesteśmy w stanie oszacować mocy testu. Znalazło to także wyraz w oprogramowaniu statystycznym, w którym podejmowane są próby szacowania tzw. empirycznej mocy testu. Skoro nie kontrolujemy prawdopodobieństwa błędu drugiego rodzaju, to przy $p > \alpha$ stajemy bezradni (nie mamy podstaw do odrzucenia hipotezy zerowej i nie mamy prawa jej przyjąć), a jeśli przyjmujemy hipotezę alternatywną przy $p < \alpha$, to jest to jedynie podjęcie decyzji o prawdziwości hipotezy alternatywnej. Nie jesteśmy jednak w stanie różnicować siły (mocy, może stopnia zaufania do podjętej decyzji) na podstawie wartości statystyki będącej podstawą testu, czy też na podstawie wartości prawdopodobieństwa oszacowanego w teście. Zrozumiała więc wydaje się próba wprowadzenia jakiejś miary, miary wielkości efektu.

6. OCENA WIELKOŚCI EFEKTU

Wielkość efektu (*effect size*) bywa także nazywana siłą zależności (*strength of association*) albo wagą badania (*treatment magnitude*).

Tytułem wprowadzenia trzy cytaty z książki A. Fielda (Field, 2009). Cytaty te ilustrują niefrasobliwość i nieodpowiedzialność w używaniu terminologii statystycznej, która w konsekwencji prowadzi do omawianego wcześniej pomieszania teorii wnioskowania statystycznego.

Pierwszy: „Proponowano wiele miar jako wielkość efektu, lecz najbardziej znane spośród nich to współczynnik *d* Cohena, współczynnik korelacji Pearsona i iloraz szans".

Drugi: „Wielkości efektu są przydatne ponieważ stanowią obiektywną miarę ważności efektu. Obojętne jakiego efektu poszukujesz, jakie zmienne zostały zmierzone i jak te zmienne były mierzone – wiemy, że współczynnik korelacji równy 0 oznacza brak efektu, a jego wartość równa 1 oznacza, że efekt jest pełny (kompletny). Cohen (1988, 1992) zaproponował szeroko wykorzystywane interpretacje, kiedy mamy do czynienia z dużym albo małym efektem:

$r = 0{,}10$ (mały efekt): w tym przypadku efekt to 1% wyjaśnionej całkowitej wariancji,

$r = 0{,}30$ (średni efekt): efekt to około 9% wyjaśnionej całkowitej wariancji,

$r = 0{,}50$ (duży (znaczny) efekt): to wyjaśnienie około 25% całkowitej wariancji".

I trzeci cytat: „Chociaż nasza statystyka *t* jest statystycznie istotna, to nie oznacza, że nasz efekt jest ważny w terminach praktycznych. By odkryć czy efekt ma znaczenie, musimy wykorzystać to, co wiemy o wielkościach efektu. Zamierzam trzymać się wielkości efektu *r*, ponieważ jest on powszechnie rozumianym, często używanym, i tak, przyznaję się, naprawdę go lubię!"

Odnośnie do pierwszego cytatu można mieć wątpliwości, czy współczynnik korelacji z próby *r* (bo tylko takim dysponujemy) jest tak popularną miarą wielkości efektu. Jeśli całkowicie zrezygnujemy z testowania hipotez, to może tak. Jeśli jednak nie chcemy całkowicie zrezygnować z testowania, a tylko je uzupełniać oceną wielkości efektu, to zależność efektów testowania współczynnika korelacji od wielkości próby powoduje, iż jego przydatność jako miernika wielkości efektu zaczyna być wątpliwa.

Zastrzeżenia dotyczące drugiego cytatu. Stwierdzenie, że nie jest ważne, jakie zmienne i w jaki sposób zostały zmierzone, oraz że nie jest ważne, jakiego efektu poszukujemy, zakrawa na statystyczną ignorancję. I nie byłoby w tym nic nagannego, gdyby nie mieszało w głowach psychologów uczących się statystyki.

Czy 1% wyjaśnionej całkowitej wariancji, gdy 99% tejże wariancji jest wyjaśniane przez nieznane nam czynniki, można uznać za mały efekt? Może raczej brak efektu. Dodatkowo, nie uwzględnia się w tej ocenie wielkości próby, na

podstawie której został on oszacowany. Wyjaśnienie 25% całkowitej wariancji to znaczny efekt przy 75% niewyjaśnionej wariancji?

I trzeci cytat. Surrealistycznie brzmi stwierdzenie, że musimy posługiwać się jakimiś miernikami, by odkryć, że efekt jest ważny z merytorycznego punktu widzenia badacza. Jest to propozycja zwalniająca z jakiegokolwiek myślenia i interpretowania wyników analizy statystycznej w terminach merytorycznych.

Książki Fielda użyłem nieprzypadkowo, gdyż jest to nagradzany podręcznik ze statystyki, w konsekwencji polecany i wykorzystywany.

Aby nie być gołosłownym jeszcze dwa głosy polemizujące z powyższymi. Pierwszy to cytowany już fragment pracy Thompsona (1994), w którym autor sugeruje, iż badacze zbyt mało rozumieją istotę testów statystycznych i zbyt mało starają się ją zrozumieć.

Z kolei Seltman łączy pojęcie wielkości efektu z mocą testu. Chcemy obliczać moc testu dla „sensownej" wielkości efektu, którą uważamy za możliwą do osiągnięcia. Podobny cel przyświeca nam przy wyborze takiej wielkości efektu, że efekt mniejszy przestałby być naukowo interesujący. Ogólnie rzecz biorąc, badacz powinien brać pod uwagę najmniejszą wielkość efektu, którą uważa za interesującą (merytorycznie, w dziedzinie badania) i próbować osiągnąć sensowną moc dla takiej wielkości efektu, dopuszczając także istnienie większej mocy dla większego efektu i mniejszej mocy dla mniejszego efektu (Seltman, 2014).

Wróćmy jeszcze na moment do podręcznika Fielda (Field, 2009, s. 332): „Przekształcając wartość *t* w wartość współczynnika korelacji r, co jest naprawdę łatwe, możemy użyć następującego równania (np. Rosenthal, 1991; Rosnow, Rosenthal, 2005):

$$r = \sqrt{\frac{t^2}{t^2 + df}}. \tag{15}$$

Skąd wziął się powyższy wzór? „Dwuwymiarowy rozkład badanych cech *X* i *Y* w populacji generalnej jest normalny lub zbliżony do normalnego. Z populacji tej wylosowano (niekoniecznie dużą) próbę *n* elementową. Na podstawie wyników tej próby oszacowano wartość współczynnika korelacji liniowej, uzyskując wartość oszacowania *r*. Przy założeniu prawdziwości hipotezy H_0: $\rho = 0$, statystyka:

$$t = \frac{r}{\sqrt{1 - r^2}}\sqrt{n - 2} \tag{16}$$

ma rozkład *t*-Studenta z $n - 2$ stopniami swobody" (Greń, 1968; Fisz, 1969).

Przekształćmy wzór (16) we wzór (15). Dla uproszczenia zapisu oznaczmy $n - 2 = df$:

$$t = \frac{r}{\sqrt{1-r^2}}\sqrt{df} \equiv t^2 = \frac{r^2 \cdot df}{1-r^2} \equiv t^2(1-r^2) = r^2 \cdot df \equiv t^2 - t^2 \cdot r^2 = r^2 \cdot df$$
$$\equiv t^2 = t^2 \cdot r^2 + r^2 \cdot df \equiv t^2 = (t^2 + df) \cdot r^2 \equiv r^2 = \frac{t^2}{t^2 + df} \equiv r = \sqrt{\frac{t^2}{t^2 + df}} \quad (17)$$

Otrzymanie wzoru (15) jest zadaniem trywialnym, ale konsekwencje tego przekształcenia będą dramatycznie poważniejsze. Zwróćmy uwagę, że statystyka *t* ma rozkład *t*-Studenta z odpowiednią liczbą stopni swobody przy założeniu normalności dwuwymiarowego rozkładu prawdopodobieństwa zmiennej (*X*, *Y*) (obie zmienne zapisane są tutaj w postaci wektora, czyli w tym przypadku w postaci zmiennej dwuwymiarowej). Co możemy powiedzieć o relacjach między *r* i *t*, gdy założenie dwuwymiarowej normalności rozkładu zmiennych (*X*, *Y*) nie będzie spełnione? Jak dalece może ono nie być spełnione (zagadnienie odporności)? Czy rozbrajające stwierdzenie, że autor po prostu lubi jakiś współczynnik, uprawnia do jego powszechnego i bezkrytycznego stosowania? Co więcej, im większa wartość współczynnika *r*, tym większa wartość statystyki *t*, a więc tym mniejsze prawdopodobieństwo odpowiadające wartości statystyki *t*. Czyli dochodzimy do relacji, im mniejsze prawdopodobieństwo w teście hipotezy:

$$\begin{cases} H_0 : \rho = 0 \\ H_1 : \rho \neq 0 \end{cases} \quad (18)$$

tym większy efekt. A takie stwierdzenie nie ma żadnych podstaw teoretycznych. W tym momencie musimy zrezygnować z testowania hipotez statystycznych.

Pomysły, które umożliwiłyby ocenę siły dowodu statystycznego pojawiły się już w latach 30. XX w. Na przykład Lindquist (1938) dyskutuje w swojej książce pojęcie „stopnia zaufania", który związany jest z odrzuceniem hipotezy bądź jej zaakceptowaniem. Od tego czasu zaproponowano wiele różnych miar oceny wielkości efektu. W niniejszym rozdziale nie będę przedstawiał ich wszystkich, ograniczę się do najbardziej intuicyjnych i takich, które są zaimplementowane w programach statystycznych, np. SPSS czy STATA.

W prezentowanych przykładach zamieszczam wydruki z pakietu SPSS. W piśmiennictwie współczynnik korelacji z próby jest oznaczany jako *r*, natomiast w programach statystycznych (SPSS, STATA, STATISTICA, SYSTAT) na oznaczenie współczynnika korelacji z próby używana jest litera *R*. Nie chciałem ingerować w wydruki z programu. Ale spowodowało to niezbyt komfortową sytuację dla Czytelnika: *r* i *R* oznaczają to samo. Brak ingerencji w wydruki komputerowe w przykładach skutkuje kolejnymi nieścisłościami. W wydrukach pojawia się prawdopodobieństwo ,000, co oznacza zaokrąglenie do trzech miejsc po przecinku obliczonego prawdopodobieństwa w teście. Nieprawdą jest, iż prawdopodobieństwo to

jest równe 0, ono jest mniejsze od 0,0005. Także kolumna zatytułowana „Istotność" może prowadzić do nieporozumień; w kolumnie tej znajduje się prawdopodobieństwo obliczone w teście, które jest porównywane z poziomem istotności testu.

6.1. Wielkość efektu w modelach regresji liniowej

W modelach regresji liniowej naturalnym miernikiem siły zależności, wielkości efektu wydaje się współczynnik determinacji z próby (R^2). Jak pamiętamy, współczynnik determinacji jest w modelach regresji liniowej kwadratem współczynnika korelacji liniowej (R w modelach jednozmiennowych) i korelacji wielokrotnej (R w modelach wielozmiennowych). Interpretacja współczynnika determinacji z próby to procent wariancji zmiennej objaśnianej wyjaśnionej przez zmienność zespołu zmiennych objaśniających. Czy wartości współczynnika determinacji określone jako: 0,25 zależność silna (duży efekt), 0,09 zależność średnia i 0,01 zależność słaba mają jednakową wymowę (wagę, znaczenie) dla modeli jednozmiennowych i wielozmiennowych?

W tym momencie pojawia się kolejne pytanie: czy rzeczywiście wartość współczynnika determinacji równa 0,25 może oznaczać zależność silną? Wartość ta oznacza, że 25% wariancji zmiennej objaśnianej jest wyjaśniane przez zmienność zmiennych objaśniających znajdujących się w modelu, ale 75% wariancji zmiennej objaśnianej jest wyjaśniane przez zmienność zmiennych, które w modelu się nie znalazły. Procent wariancji wyjaśnionej jest nieporównywalnie mniejszy od części niewyjaśnionej przez zmienne w modelu. Wydaje się, że twórcy przedziałów dla współczynnika determinacji zasugerowali się wartością współczynnika korelacji liniowej. Wartości współczynnika determinacji 0,25 odpowiada wartość współczynnika korelacji liniowej 0,5. Warto też pamiętać o uwagach Thompsona (1994): „Jakie są efekty wynikające z wielkości próbki?" I nie da się określać przedziałów wielkości efektów bez uwzględnienia wielkości próby. Wartość współczynnika korelacji 0,9 dla próby trzyelementowej jest nic nieznacząca, a 0,5 dla próby 100 elementowej niesie już sporo informacji.

Valentine i Cooper (2003) zauważają, że zaproponowane przez Cohena (1988) „punkty odcięcia" dla współczynnika korelacji 0,1; 0,3 i 0,5 są odzwierciedleniem typowej wielkości efektu, z jaką można się spotkać w naukach behawioralnych jako całości. Cohen przestrzegał jednak przed używaniem tych granic do interpretowania relacji polegającej na ocenie wagi zagadnienia czy problemu w obrębie poszczególnych dyscyplin nauk społecznych albo obszarów tematycznych. Pewne obszary, jak np. edukacja, prawdopodobnie mają mniejsze wielkości efektów niż inne, zatem dosłowne stosowanie granic Cohena może wprowadzać w błąd. Ponieważ granice wielkości efektu Cohena pozwalają tylko na najogólniejszą interpretację miary wielkości efektu, powinny więc one być wykorzystywane z dużą ostrożnością. Ich najpoważniejszą ułomnością jest to, iż

w większości przypadków proporcja wyjaśnionej wariancji nie powinna być używana jako wielkość efektu. To ostatnie stwierdzenie dotyczy sytuacji innych niż modelowanie zależności metodami regresji liniowej.

W książce Cohena (1988) proponowany jest także inny miernik wielkości efektu w wielozmiennowych modelach regresji liniowej, f^2, ale jest on prostą funkcją współczynnika determinacji, przedziały dla f^2 są pochodnymi granic przedziałów dla R^2 i miernik ten nic nowego do oceny wielkości efektu nie wnosi.

Skoro próbujemy oceniać wielkość efektu dla wielozmiennowego modelu regresji liniowej, może warto by pokusić się o ocenę wielkości efektu związanego z każdą ze zmiennych umieszczoną w modelu. W tych modelach dysponujemy standaryzowanymi współczynnikami regresji, ale pozwalają one jedynie na porangowanie zmiennych objaśniających pod względem siły zależności ze zmienną objaśnianą. Miernik „zmiana R^2" jest mało przydatny z powodu przyjętych granic dla oceny wielkości efektu. W tym sensie najczęściej tylko pierwsza zmienna wprowadzana do modelu powoduje stosunkowo duży przyrost współczynnika determinacji, a kolejne będą traktowane jako mające zależność mniejszą niż słabą.

Przeanalizujmy przykład 2. We wszystkich przykładach w tym rozdziale będą wykorzystane wyniki badania Bohdana Dudka i jego zespołu nad wpływem stresu zawodowego na stan zdrowia (Dudek, 2007).

Przykład 2. W modelu regresyjnym sugeruję istnienie liniowej zależności między zmienną „subiekt" (subiektywne odczucie stresu związanego z pracą) i zmiennymi objaśniającymi: „SOC" (poczucie koherencji), „GHQ_suma" (subiektywna ocena stanu zdrowia według 28-pytaniowego kwestionariusza Goldberga) oraz zmiennymi opisującymi nastrój: „wrogość", „zakłopotanie", „przygnębienie", „znużenie", „życzliwość", „napięcie" i „wigor". Użyłem krokowej metody budowy modelu z prawdopodobieństwem wprowadzenia zmiennej równym 0,05 i usunięcia zmiennej 0,051.

Końcowy model powstał po ośmiu krokach. W pierwszym kroku wprowadzona została do modelu zmienna „przygnębienie", jednakże w piątym kroku została ona usunięta z modelu i w ósmym kroku została usunięta z modelu zmienna „napięcie", wprowadzona w kroku trzecim (tab. 1).

Tabela 1. Model – podsumowanie

Model	*R*	*R*-kwadrat	Skorygowane *R*-kwadrat	Błąd standardowy oszacowania	Statystyki zmiany				
					zmiana *R*-kwadrat	*F* zmiany	*df*1	*df*2	istotność *F* zmiany
1	,590[a]	,348	,346	24,4212	,348	233,653	1	438	,000
2	,638[b]	,407	,405	23,3084	,059	43,820	1	437	,000
3	,654[c]	,428	,424	22,9180	,021	16,016	1	436	,000
4	**,661[d]**	**,436**	**,431**	**22,7812**	**,008**	**6,250**	**1**	**435**	**,013**
5	,658[e]	,433	,429	22,8175	-,003	2,390	1	435	,123

Tab. 1 (cd.)

Model	*R*	*R*-kwadrat	Skorygowane *R*-kwadrat	Błąd standardowy oszacowania	Statystyki zmiany				
					zmiana *R*-kwadrat	*F* zmiany	*df*1	*df*2	istotność *F* zmiany
6	,666[f]	,443	,438	22,6402	,010	7,856	1	435	,005
7	**,672[g]**	**,451**	**,445**	**22,5077**	**,008**	**6,136**	**1**	**434**	**,014**
8	,670[h]	,449	,444	22,5270	-,002	1,745	1	434	,187

[a] Predyktory: (Stała), przygneb. [b] Predyktory: (Stała), przygneb, SOC. [c] Predyktory: (Stała), przygneb, SOC, napiecie. [d] **Predyktory: (Stała), przygneb, SOC, napiecie, zaklopot**. [e] Predyktory: (Stała), SOC, napiecie, zaklopot. [f] Predyktory: (Stała), SOC, napiecie, zaklopot, wrogość. [g] **Predyktory: (Stała), SOC, napiecie, zaklopot, wrogosc, GHQ_suma**. [h] Predyktory: (Stała), SOC, zaklopot, wrogosc, GHQ_suma.

Źródło: opracowanie własne.

W kroku czwartym wprowadzona została do modelu zmienna „zakłopotanie", dla której zmiana R^2, czyli przyrost współczynnika determinacji jest równy 0,008 – oznacza to zależność słabszą niż słaba. Analogicznie wprowadzona w kroku siódmym zmienna „GHQ_suma" zwiększa współczynnik determinacji też tylko o 0,008. Oczywiście z punktu widzenia teorii Neymana–Pearsona testowania hipotez oba te przyrosty R^2 są istotnie różne od zera (prawdopodobieństwo w odpowiednich testach jest mniejsze od 0,05).

Z kolei przyjrzyjmy się standaryzowanym współczynnikom regresji (β – *beta*). W poniższej tabelce pokazuję tylko ostatni, końcowy model (tab. 2).

Tabela 2. Współczynniki

Współczynniki[a]								
Model		Współczynniki niestandaryzowane		Współczynniki standaryzowane	*t*	Istotność	95,0% przedział ufności dla *B*	
		B	błąd standardowy	*beta*			dolna granica	górna granica
8	(Stała)	138,257	11,827		11,690	,000	115,012	161,502
	SOC	–,335	,065	–,259	–5,135	,000	–,464	–,207
	zaklopot	1,488	,445	,190	3,344	,001	,614	2,363
	wrogosc	,750	,216	,193	3,473	,001	,326	1,175
	GHQ_ suma	,429	,146	,148	2,942	,003	,142	,716

[a] Zmienna zależna: subiekt.

Źródło: opracowanie własne.

Mimo że przyrosty R^2 zmiennych „zakłopotanie" i „GHQ_suma" były jednakowe i wynosiły 0,008, to standaryzowane współczynniki regresji dla tych zmiennych różnią się; dla zmiennej „zakłopotanie" jest to 0,190, dla zmiennej „GHQ_ suma" 0,148. Niby jest to sensowne, gdyż każda następna zmienna wprowadzana

do modelu mniej do niego wnosi w zakresie wyjaśniania wariancji zmiennej objaśnianej, ale w tym konkretnym przypadku to się nie sprawdziło. Jeśli chodzi o zmienną „wrogość", to jej wprowadzenie do modelu zwiększyło R^2 o 0,010, a więc nieznacznie więcej niż zmiennych „zakłopotanie" oraz „GHQ_suma" i *beta* zmiennej „wrogość" jest nieznacznie większa: 0,193. Ale czy możemy mówić tutaj o jakiejkolwiek ocenie wielkości efektu? Podobne zastrzeżenia do wykorzystywania standaryzowanych współczynników regresji jako miar wielkości efektu ma Greenland i in. (1986, 1991). Mimo że ich obiekcje dotyczą tego typu mierników wielkości efektów w zagadnieniach biologicznych oraz zdrowia publicznego, to istota problemu jest taka sama. I czy potrzebne jest wprowadzanie jeszcze jednego sztucznego miernika? Sądzę, że znacznie ważniejsze od różnych mierników jest przeprowadzenie przez badacza głębokiej, rzetelnej, merytorycznej analizy uzyskanych wyników modelowania statystycznego.

6.2. Wielkość efektu w modelach regresji logistycznej

W modelach regresji logistycznej spotykamy się z takimi samymi problemami związanymi z oceną wielkości efektu, jak w modelach regresji liniowej, z całościową oceną wielkości efektu modelu oraz oceną wielkości efektu poszczególnych zmiennych modelu.

Dla modeli regresji logistycznej podejmowane były próby skonstruowania miernika podobnego do współczynnika determinacji (R^2) w regresji liniowej. Powstało kilka różnych tzw. pseudo-R^2, z których żaden nie ma właściwości współczynnika R^2 z modelu liniowego. W pakiecie SPSS są zaimplementowane dwa pseudo R^2, jest to współczynnik R^2 Nagelkerke'a oraz współczynnik R^2 Coxa i Snella (Nagelkerke, 1991).

Przyjmijmy następujące oznaczenia, aby móc przedstawić wzory dla R^2 zaimplementowanych w SPSS i najpopularniejszego pseudo-R^2, czyli $\mathbf{R}^2_{\mathrm{L}}$: L_F jest wartością funkcji wiarygodności modelu zawierającego wszystkie predyktory (model pełny, końcowy model w konkretnym badaniu); L_0 jest wartością funkcji wiarygodności modelu zawierającego tylko stałą, n oznacza ogólną liczebność próbki. Współczynnik Coxa i Snella wyrażony jest wówczas wzorem:

$$\mathbf{R^2_{CS} = 1 - \left(\frac{ln(L_0)}{ln(L_F)} \right)^{\frac{2}{n}}} \tag{19}$$

zaś współczynnik Nagelkerke'a wzorem:

$$\mathbf{R^2_N = \frac{1 - [ln(L_0) / ln(L_F)]^{2/n}}{1 - [ln(L_0)]^{2/n}}} \tag{20}$$

Różne „mutacje" pseudo-R^2, zarówno te przedstawione powyżej, jak i inne, omawiane są w pracach: Magee (1990), Allen i Le (2007), Agresti (1990), Hilbe (2009).

Najpopularniejszy, najczęściej używany miernik pseudo-R^2 jest zdefiniowany jako (Hilbe, 2009; Menard, 2000; Hosmer, Lemeshow, 1989):

$$\mathbf{R_L^2 = 1 - \frac{ln(L_F)}{ln(L_0)}} \tag{21}$$

Menard (2000) napisał, że „po pierwsze i najważniejsze, $\mathbf{R_L^2}$ ma najbardziej intuicyjnie uzasadnioną interpretację jako proporcjonalna redukcja miary błędu porównywalna z klasycznym R^2". Jednakże stosując $\mathbf{R_L^2}$ nie jesteśmy pewni wpływu predyktorów (czynników ryzyka) na rezultat. Co na przykład oznacza $\mathbf{R_L^2 = 0{,}10}$ w terminach zmiany prawdopodobieństwa albo szansy? Nikt na to sensownie nie potrafił odpowiedzieć. Co więcej, praktycznie żaden z pseudo-R^2 nie może być wykorzystywany jako miernik dobroci dopasowania modelu do danych empirycznych, czego konsekwencją jest brak miernika wielkości efektu modelu regresji logistycznej jako całości.

Z kolei przyjrzyjmy się miernikom wielkości efektu dla pojedynczych czynników ryzyka (predyktorów) w modelu regresji logistycznej. Tabachnick i Fidell (2007) sugerują, powołując się na pracę Chinn (2000), iż można przekształcić iloraz szans do współczynnika Cohena *d*, który z kolei może być przekształcony w η^2:

$$\mathbf{d = ln(OR) / 1{,}81}$$
$$\mathbf{\eta^2 = \frac{d^2}{d^2 + 4}} \tag{22}$$

Pomijając magiczne działania dzielenia przez 1,81, zaproponowany sposób oceny wielkości efektu znajduje zastosowanie tylko w przypadku ciągłych czynników ryzyka. Dla dyskretnych czynników ryzyka otrzymujemy ilorazy szans (ORs) dla każdej wartości tegoż czynnika w odniesieniu do przyjętej kategorii odniesienia. Co nam da przeliczenie otrzymanych ilorazów szans dla poszczególnych kategorii czynników ryzyka do wartości η^2 w odniesieniu do całej zmiennej (czynnika ryzyka)? Nie znajduję odpowiedzi na takie pytanie.

6.3. Wielkość efektu w modelach analizy wariancji

W podręczniku Tabachnick i Fidell (2007) znajdujemy trzy mierniki wielkości efektu wykorzystywane w modelach analizy wariancji. Jest to współczynnik η^2 wyrażony wzorem:

$$\eta^2 = \frac{SS_{effect}}{SS_{total}} \tag{23}$$

cząstkowy współczynnik η^2:

$$\eta_p^2 = \frac{SS_{effect}}{SS_{effect} + SS_{error}} \tag{24}$$

i współczynnik $\hat{\omega}^2$:

$$\hat{\omega}^2 = \frac{SS_{effect} - (df_{effect}) \cdot MS_{error}}{SS_{total} + MS_{error}}. \tag{25}$$

SS_{effect} – statystyka ta mierzy stopień, w jakim średnie podgrup wyznaczonych przez poziomy czynnika różnią się od ogólnej średniej,

SS_{total} to ogólna suma kwadratów (w SPSS oznaczana jako suma kwadratów ogółem), suma kwadratów odchyleń każdej obserwacji w eksperymencie od ogólnej średniej,

SS_{error} oznacza zmienność spowodowaną błędem eksperymentalnym, to suma kwadratów związana z każdym pojedynczym efektem (czynnikiem albo efektem interakcyjnym); w modelu analizy wariancji; interpretowana bywa jako łączna miara zmienności obserwacji wewnątrz grup wyznaczonych przez poziomy czynnika,

MS_{error} to średnia SS_{error} : $MS_{error} = SS_{error}/df_{error}$,

df – oznacza liczbę stopni swobody odpowiedniej statystyki.

Miernik $\hat{\omega}^2$ jest ograniczony do oceny efektów międzyobiektowych w planach analizy wariancji z równymi liczebnościami w komórkach, czyli jest przydatniejszy dla planów z powtarzanymi obserwacjami. Problem z η^2 polega na tym, że wielkość tego miernika dla każdego poszczególnego efektu zależy w pewnym stopniu od znaczenia i liczby innych efektów w planie badawczym (Tabachnick, Fidell, 2007). Skutek występowania w planie badawczym większej liczby efektów minimalizuje miernik cząstkowej η^2. **Uwaga**: mierniki $\boldsymbol{\eta^2}$ **i** $\boldsymbol{\eta_p^2}$ w jednoczynnikowej analizie wariancji są jednakowe. W innych modelach analizy wariancji $\boldsymbol{\eta_p^2 < \eta^2}$, co wynika z porównania wzorów (23) i (24).

Wróćmy na moment do programu SPSS. Program nie oblicza wielkości efektu w jednoczynnikowej analizie wariancji. Ale ponieważ podawane są odpowiednie sumy kwadratów, można to zrobić samodzielnie. W analizach wieloczynnikowych obliczane są cząstkowe η^2, zaś wartość miernika η^2 można policzyć, korzystając z odpowiednich sum kwadratów. W modelach analizy wariancji z powtarzanymi pomiarami też liczone są cząstkowe η^2.

Oczywiście wartości mierników można obliczyć, ale co z nich wynika? W podręczniku (Tabachnick, Fidell, 2007) podane są za Cohenem (1988) przedziały dla η^2. Efekt słaby to $\eta^2 = 0{,}01$, efekt umiarkowany to $\eta^2 = 0{,}09$ i efekt duży to wartość $\eta^2 = 0{,}25$. Sink i Mvududu (2010) proponują nieco inne granice dla η^2, mianowicie efekt słaby to $\eta^2 = 0{,}01$, umiarkowany $\eta^2 = 0{,}06$, a silny to $\eta^2 = 0{,}14$. Zauważają oni, że wartości progowe dla cząstkowej η^2 są zwykle mniejsze niż te dla η^2; stąd granice dla oceny efektu jako słabego, umiarkowanego i silnego dla $\boldsymbol{\eta}_p^2$ są prawdopodobnie zbyt duże, zatem muszą być interpretowane bardzo ostrożnie. Warto zwrócić uwagę, że przedziały zaproponowane przez Sinka i Mvududu są niższe niż zaproponowane przez Cohena. Może przyjęcie granic Cohena dla η^2, zaś propozycji Sinka i Mvududu jako granic dla $\boldsymbol{\eta}_p^2$ byłoby sensownym rozwiązaniem, ale nigdzie nie znalazłem takiej propozycji.

W programie SPSS cząstkowe η^2 są także obliczane w modelach analizy kowariancji.

6.4. Porównywanie dwóch wartości oczekiwanych

Porównując dwie wartości oczekiwane w sytuacji równych wariancji w obu grupach, Cohen (1988) zaproponował miernik wielkości efektu w postaci:

$$\mathbf{d} = \frac{\mu_1 - \mu_2}{\sigma}. \tag{26}$$

We wzorze tym występują symbole oznaczające prawdziwe, a więc nieznane nam, wartości parametrów: wartości oczekiwane i odchylenie standardowe. Oczywiście w praktyce będziemy wykorzystywali wartości estymatorów odpowiednich parametrów, i to dla nieco ogólniejszej sytuacji, tj. niejednorodnych wariancji w porównywanych grupach:

$$\mathbf{d} = \frac{\bar{x}_1 - \bar{x}_2}{s_{wspólne}}; \qquad s_{wspólne} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2}} \tag{27}$$

$$\mathbf{d} = \frac{\bar{x}_1 - \bar{x}_2}{s_{wspólne}}; \qquad s_{wspólne} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \tag{28}$$

gdzie n_1 i n_2 to liczebności próbek, na podstawie których obliczane były średnie i wariancje z próby. Wzór (27) pochodzi z opracowania Thalheimera i Cook (2002), zaś wzór (28) z pracy Volkera (2006).

Dla tak określonego miernika wielkości efektu Cohen (1988) zaproponował następujące granice: $d = 0{,}2$ oznacza efekt słaby, $d = 0{,}5$ efekt średni, zaś $d = 0{,}8$ efekt silny. W uzasadnieniu tych granic Cohen wykorzystał normalność rozkładu

Tabela 3. Statystyki opisowe

Wyszczególnienie		N	Średnia	Odchylenie standardowe	Błąd standardowy	95% przedział ufności dla średniej		Minimum	Maksimum
						dolna granica	górna granica		
cholest	,0	303	199,871	36,4017	2,0912	195,756	203,987	118,0	365,0
	1,0	135	212,867	35,8394	3,0846	206,766	218,967	123,0	340,0
	ogółem	438	203,877	36,6836	1,7528	200,432	207,322	118,0	365,0
HDL	,0	303	61,547	13,0048	,7471	60,077	63,017	33,7	107,4
	1,0	135	58,470	12,5301	1,0784	56,337	60,603	26,0	93,0
	ogółem	438	60,599	12,9245	,6176	59,385	61,812	26,0	107,4
cukier	,0	303	84,917	11,2702	,6475	83,643	86,192	62,0	157,0
	1,0	135	87,230	13,3450	1,1486	84,958	89,501	54,0	129,0
	ogółem	438	85,630	11,9804	,5724	84,505	86,755	54,0	157,0
skurcz	,0	305	129,502	9,4029	,5384	128,442	130,561	105,0	160,0
	1,0	132	141,629	17,2376	1,5003	138,661	144,597	105,0	190,0
	ogółem	437	133,165	13,4907	,6454	131,896	134,433	105,0	190,0

Źródło: opracowanie własne.

badanej cechy w porównywanych grupach, co osłabia argumentację. Czy argumentacja ta byłaby równie skuteczna, gdy rozkłady badanej cechy nie będą normalne? W praktyce znaczniej częściej mamy do czynienia z cechami o rozkładach niebędących normalnymi niż z rozkładami normalnymi.

Przykład 3. Wróćmy do wyników badania Bohdana Dudka i jego zespołu nad wpływem stresu zawodowego na stan zdrowia pracowników wybranych służb mundurowych (Dudek, 2007). Porównamy wartości oczekiwane cholesterolu całkowitego (zmienna: „cholest"), frakcji HDL cholesterolu (zmienna: „HDL"), poziomu cukru na czczo (zmienna: „cukier"), oraz ciśnienia skurczowego (zmienna: „skurcz") w grupach określonych przez wartości zmiennej „ukl_kraz" (zob. tab. 3–7). Zmienna „ukl_kraz" jest zmienną dwustanową:

$$\begin{cases} H_0: \text{zmienne X i Y są niezależne} \\ H_1: \text{zmienne X i Y nie są niezależne} \end{cases}$$

Tabela 4. Test jednorodności wariancji

Wyszczególnienie	Test Levene'a	*df*1	*df*2	Istotność
cholest	,011	1	436	,915
HDL	,073	1	436	,787
cukier	4,399	1	436	,037
skurcz	80,291	1	435	,000

Źródło: opracowanie własne.

Tabela 5. Jednoczynnikowa ANOVA

Wyszczególnienie		Suma kwadratów	*df*	Średni kwadrat	*F*	Istotność
cholest	między grupami	15771,762	1	15771,762	12,016	,001
	wewnątrz grup	572293,580	436	1312,600		
	ogółem	588065,342	437			
HDL	między grupami	883,923	1	883,923	5,344	,021
	wewnątrz grup	72113,956	436	165,399		
	ogółem	72997,879	437			

Źródło: opracowanie własne.

Dla dwóch badanych cech: cholesterol całkowity („cholest") i „HDL" nie możemy odrzucić hipotez o jednorodności wariancji, natomiast dla zmiennych: „cukier" i ciśnienie skurczowe („skurcz") wariancje w porównywanych grupach

są różne. Wynikają z tego oczywiste i znane konsekwencje; dla zmiennych: cholesterol całkowity („cholest") i „HDL" zastosujemy klasyczny test *F*-Snedecora do porównywania wartości oczekiwanych, zaś dla zmiennych: „cukier" i „skurcz" testy Welcha i Browna-Forsythe'a, które są odporne na niespełnianie tego założenia.

Tabela 6. Mocne testy równości średnich

Wyszczególnienie		Statystyka[a]	*df*1	*df*2	Istotność
cukier	Welch	3,075	1	222,715	,081
	Brown-Forsythe	3,075	1	222,715	,081
skurcz	Welch	57,880	1	165,728	,000
	Brown-Forsythe	57,880	1	165,728	,000

[a] Rozkład *F* asymptotyczny.
Źródło: opracowanie własne.

Wartości oczekiwane zmiennych: cholesterol całkowity, HDL i ciśnienie skurczowe różnią się w porównywanych grupach: osób wolnych od chorób układu krążenia i osób ze zdiagnozowaną chorobą układu krążenia. Natomiast dla zmiennej „cukier" nie mamy podstaw do odrzucenia hipotezy o równości wartości oczekiwanych w porównywanych grupach.

Tabela 7. Obliczenie miernika wielkości efektu *d* Cohena dla analizowanych w przykładzie zmiennych

Badana zmienna	Ukl_kraz	Średnia	SD	$\bar{x}_1 - \bar{x}_2$	$s_{wspólne}$	*d* Cohena
cholest	0 1	199,87 212,87	36,40 35,84	–13,00	36,23	–0,36
HDL	0 1	61,55 58,47	13,00 12,53	3,08	12,86	0,24
cukier	0 1	84,92 87,23	11,27 13,34	–2,31	11,94	–0,19
ciśnienie skurczowe	0 1	129,50 141,63	9,40 17,24	–12,13	12,31	–0,99

Źródło: opracowanie własne.

Znak miernika *d* Cohena nie jest ważny dla wielkości efektu, więc nie będę się nim zajmował. Ograniczę się do dyskusji mierników dla dwóch zmiennych, mianowicie dla „HDL" i „cukru" we krwi. Wyniki testu hipotezy o równości wartości oczekiwanych dla zmiennej „HDL" doprowadziły do podjęcia decyzji, iż wartości oczekiwane w grupach są różne. Decyzja ta dotyczy wartości

oczekiwanych, czyli parametrów teoretycznych, traktowanych jako prawdziwe wartości parametrów, gdyż decyzja wynikająca z testowania hipotez statystycznych dotyczy populacji generalnej. Wartość miernika *d* Cohena jest obliczana na podstawie próby. Jego wartość 0,24 świadczy o tym, że zależność jest nieznacznie większa niż słaba. Jednakże testowanie hipotez i obliczanie wielkości efektów odbywają się w różnych przestrzeniach. Ile mają ze sobą wspólnego? Nie należy także zapominać o merytorycznej ocenie wielkości różnicy „HDL" w porównywanych grupach pracowników służb mundurowych. Czy różnica między średnimi „HDL" w porównywanych grupach wielkości 3 jednostek, przy rozrzucie mierzonym odchyleniem standardowym wielokrotnie większym niż uzyskana różnica, ma jakiekolwiek znaczenie z lekarskiego punktu widzenia? To że otrzymaliśmy jakiś wynik z obliczeń podczas analizy statystycznej nie jest argumentem rozstrzygającym, jest zaledwie jedną z przesłanek do podjęcia decyzji.

Drugim parametrem, który już teraz krótko omówię, jest poziom cukru w surowicy krwi na czczo. W wyniku testowania hipotezy o równości wartości oczekiwanych w porównywanych grupach nie mieliśmy podstaw do odrzucenia hipotezy zerowej. Oczywiście niemożność odrzucenia hipotezy zerowej stawia nas w bardzo trudnym położeniu – jesteśmy zawieszeni, nie mogąc podjąć żadnej decyzji. Czy obliczenie wielkości efektu $d = 0{,}19$ (zależność słaba, ale zależność istnieje) ułatwia nam odnalezienie się w tej sytuacji? Sądzę, że nie. Znów ważniejsza wydaje mi się merytoryczna ocena różnicy średnich w próbie. Czy różnica 2,31 przy średniej wartości około osiemdziesięciu kilku jednostek i odchyleniach standardowych 11 i 13 dostarcza jakiejś istotnej informacji, na przykład lekarzowi?

6.5. Ocena niezależności dwóch zmiennych dyskretnych

Podstawowymi testami statystycznymi do oceny niezależności dwóch zmiennych dyskretnych są: test *chi*-kwadrat niezależności, nazywanym także testem *chi*-kwadrat Pearsona oraz dokładny test Fishera. O tym, który z nich będzie zastosowany decydują liczebności oczekiwane w komórkach tablicy liczebności, nazywanej też tablicą kontyngencji. Efektem testowania hipotez:

$$\begin{cases} H_0: \text{zmienne X i Y są niezależne} \\ H_1: \text{zmienne X i Y nie są niezależne} \end{cases} \tag{29}$$

jest podjęcie decyzji o zależności zmiennych X i Y albo decyzja o braku możliwości odrzucenia hipotezy o ich niezależności. Jeśli podjęliśmy decyzję o zależności zmiennych X i Y (podjęliśmy decyzję, iż nie są one niezależne), mamy do dyspozycji całą gamę różnego rodzaju mierników siły zależności między badanymi

zmiennymi. Nie będę ich tutaj wymieniał, odsyłam Czytelnika do podręcznika (Szymczak, 2010).

Cohen (1988) zaproponował jako miernik wielkości efektu:

$$\Phi = \sqrt{\frac{\chi^2}{n}} = w \tag{30}$$

czyli jeden z symetrycznych mierników siły zależności między dwiema zmiennymi dyskretnymi wykorzystujących wartość statystyki *chi*-kwadrat. W powyższym wzorze χ^2 oznacza wartość statystyki *chi*-kwadrat, będącej podstawą testu *chi*-kwadrat niezależności, a *n* to suma liczebności we wszystkich komórkach tablicy kontyngencji. Cohen określił dla tego miernika pewne wartości, które mają charakteryzować wielkość efektu. I tak, $w = 0{,}10$ oznacza małą wielkość efektu, $w = 0{,}30$ to średnia wielkość efektu, zaś $w = 0{,}50$ to duża wielkość efektu.

W odniesieniu do miernika wielkości efektu w pojawiają się dwa główne zastrzeżenia. Pierwsze: czy miernik w mierzy jakąś uniwersalną zależność, czy może tylko jakiś szczególny, specyficzny typ zależności? Żadna pojedyncza miara nie jest najlepsza we wszystkich sytuacjach. Dla tablic $r \times c$ rzadko jest możliwe satysfakcjonujące określenie stopnia zależności między zmiennymi za pomocą wartości jednego miernika. Drugie zastrzeżenie: na ile poprawny jest miernik w, gdy do testowania niezależności dwóch zmiennych dyskretnych, ze względu na niewielkie liczebności oczekiwane w komórkach tablicy kontyngencji, powinniśmy zastosować dokładny test Fishera?

Przykład 4. Rozważmy teraz pewną hipotetyczną sytuację w grupie kobiet (w badanej próbie jest ich tylko 24). Oceńmy zależność między zmiennymi „grupa": 1 → strażacy, 2 → pracownicy służb więziennych, 3 → policjanci oraz zmienną „hobby": 1 → osoba znajduje czas na uprawianie swojego hobby, 2 → nie znajduje czasu na hobby.

Tabela 8. Tabela krzyżowa hobby × grupa

Wyszczególnienie			Grupa			Ogółem
			1,0	2,0	3,0	
Hobby	1,0	liczebność	0	7	8	15
		% z grupa	0,0%	63,6%	72,7%	62,5%
	2,0	liczebność	2	4	3	9
		% z grupa	100,0%	36,4%	27,3%	37,5%
Ogółem		liczebność	2	11	11	24
		% z grupa	100,0%	100,0%	100,0%	100,0%

Źródło: opracowanie własne.

Tabela 9. Testy *chi*-kwadrat

Wyszczególnienie	Wartość	*df*	Istotność asymptotyczna (dwustronna)	Istotność dokładna (dwustronna)	Istotność dokładna (jednostronna)	Estymacja punktowa prawdopodobieństwa
Chi-kwadrat Pearsona	3,830[a]	2	,147	,173		
Iloraz wiarygodności	4,443	2	,108	,173		
Dokładny test Fishera	3,272			,250		
Test związku liniowego	2,396[b]	1	,122	,191	,111	,082
N ważnych obserwacji	24					

[a] 66,7% komórek (4) ma liczebność oczekiwaną mniejszą niż 5. Minimalna liczebność oczekiwana wynosi ,75. [b] Wartość standaryzowana wynosi –1,548.

Źródło: opracowanie własne.

Tabela 10. Miary kierunkowe

Wyszczególnienie			Wartość	Asymptotyczny błąd standardowy[a]	Przybliżone *T*[b]	Istotność przybliżona	Istotność dokładna
Nominalna przez nominalna	*lambda*	symetryczna	,136	,172	,736	,462	
		zmienna zależna: hobby	,222	,139	1,477	,140	
		zmienna zależna: grupa	,077	,286	,259	,796	
	tau Goodmana i Kruskala	zmienna zależna: hobby	,160	,067		,160[c]	,173
		zmienna zależna: grupa	,037	,044		,429[c]	,367
	współczynnik niepewności	symetryczna	,117	,072	1,494	,108[d]	,173
		zmienna zależna: hobby	,140	,094	1,494	,108[d]	,173
		zmienna zależna: grupa	,100	,058	1,494	,108[d]	,173

[a] Nie zakładając hipotezy zerowej. [b] Użyto asymptotycznego błędu standardowego, przy założeniu hipotezy zerowej. [c] W oparciu o aproksymację rozkładu *chi*-kwadrat. [d] Prawdopodobieństwo testowe ilorazu wiarygodności *chi*-kwadrat.

Źródło: opracowanie własne.

Tabela 11. Miary symetryczne

Wyszczególnienie		Wartość	Istotność przybliżona	Istotność dokładna
Nominalna przez nominalna	*phi*	,399	,147	,173
	V Kramera	,399	,147	,173
	współczynnik kontyngencji	,371	,147	,173
N ważnych obserwacji		24		

Źródło: opracowanie własne.

Pod tabelką „Testy *chi*-kwadrat" (tab. 9) mamy komunikat, iż 66,7% komórek (4) ma liczebność oczekiwaną mniejszą niż 5. Oznacza to, że nie powinniśmy stosować testu *chi*-kwadrat niezależności. A wartość miernika wielkości efektu obliczona na podstawie wartości statystyki *chi*-kwadrat rzeczywiście jest równa 0,399:

$$\Phi = w = \sqrt{\frac{3,830}{24}} = 0,399$$

Jak zauważyliśmy przed momentem, nie powinniśmy stosować testu *chi*-kwadrat niezależności, a więc nie powinnyśmy wykorzystywać do obliczania wielkości efektu wartości statystyki *chi*-kwadrat. Wartość statystyki będącej podstawą dokładnego testu Fishera ma inną wartość i analogiczne obliczenia jak dla *chi*-kwadrat dadzą inny wynik:

$$\sqrt{\frac{3,272}{24}} = 0,369$$

To, że wielkości efektu obliczone na podstawie testu *chi*-kwadrat i dokładnego testu Fishera nie różnią się zbyt mocno, nie jest żadnym argumentem. Jak wielokrotnie zauważałem, wartość statystyki nie jest argumentem za jej stosowaniem. Dodatkowo, zbieżność wyników jest rezultatem zastosowania asymptotycznego rozwiązania w teście dokładnym Fishera, stosowanego dla dużych prób. A dokładny test Fishera jest głównie stosowany dla małych prób.

Zwróćmy jeszcze uwagę na prawdopodobieństwa w obu omawianych testach. W teście *chi*-kwadrat niezależności prawdopodobieństwo jest równe 0,173, w dokładnym teście Fishera 0,250. Oba te prawdopodobieństwa są większe od zazwyczaj przyjmowanego poziomu istotności $\alpha = 0,05$. Zatem w obu przypadkach (pamiętajmy jednak, że jedno z rozwiązań jest nieprawidłowe) nie mamy podstaw do odrzucenia hipotezy zerowej, iż badane zmienne są niezależne. W świetle zaproponowanych przez Cohena przedziałów dla wielkości miernika w efekt jest między małym a średnim.

Czy decyzje podjęte na podstawie testowania hipotez i szacowania wielkości efektu są sprzeczne? Nie można na to jednoznacznie odpowiedzieć, gdyż brak podstaw do odrzucenia hipotezy zerowej o niezależności badanych zmiennych dyskretnych praktycznie nie jest żadną decyzją. Gdyby udało się oprzeć szacowanie wielkości efektu na podstawach teoretycznych, być może można by zbudować nowy paradygmat statystyki. Przez 25 lat od propozycji Cohena takie uwarunkowania teoretyczne nie pojawiły się. Próby budowy nowego paradygmatu idą raczej w kierunku wykorzystania pojęcia wiarygodności (Blume, 2002; Royall, 1997, 2000), ewentualnie rozwiązań bayesowskich.

7. PODSUMOWANIE

W artykule tym zwróciłem uwagę na pewne słabości teorii testowania hipotez statystycznych, jak również na konsekwencje wymieszania dwóch teorii: Fishera i Neymana–Pearsona. Najpoważniejszą konsekwencją obecnego paradygmatu statystyki wydaje mi się „mała precyzja czy mała delikatność" obecnie używanych metod. Skutkuje to próbami konstruowania pewnych mierników, które miałyby „doprecyzowywać" wnioski uzyskane z testowania hipotez. Sądzę, że niektórzy badacze stosujący dodatkowe mierniki idą za daleko, próbując zastępować testowanie hipotez szacowaniem wielkości efektu. W niektórych z rozważanych wyżej przykładów widać pewne sprzeczności między wynikiem testowania a wnioskowaniem na podstawie oszacowania wielkości efektów. Argumenty za szacowaniem wielkości efektów są różne, m.in. takie, że wielkość efektu powinna być szacowana na mocy autorytetu instytucjonalnego (Volker, 2006). Innego rodzaju argumenty za szacowaniem wielkości efektu to umożliwienie wyjścia poza konkluzje statystyczne (Volker, 2006; Kline, 2013). Istnieją też argumenty przeciwko ocenie wielkości efektów (Denis, 2003). Argumenty, na które powołuje się Denis pochodzą od innych autorów.

Przykładowo, Favreau (1997) uważa, że ograniczeniem wykorzystania wielkości efektu jest jego zależność od operacjonalizacji zmiennej zależnej. Z kolei Dooling i Danks (1975) stwierdzają, że psychologia, z powodu swojej natury wykorzystującej plany eksperymentalne, po prostu nie jest gotowa do rozpoczęcia adekwatnej interpretacji statystyki wielkości efektu. Natomiast nikt nie podnosi problemu braku podstaw teoretycznych dla interpretacji wielkości efektu. Sam Denis uważa, że ocena wielkości efektów przynosi więcej korzyści niż strat.

W piśmiennictwie pojawia się jeszcze jedno, ogromnie ważne pytanie. Czy ocena wielkości efektu ma stanowić uzupełnienie testowania hipotez, czy też ma je całkowicie zastąpić? Na szczęście w tej materii większość autorów uważa, że ocena wielkości efektu to ważne, ale tylko uzupełnienie testowania hipotez. Na przykład Chow (1996) stwierdza: „Orędownicy wielkości efektu są gotowi do

stosowania niestatystycznych kryteriów nawet wówczas, gdy jest niemożliwym wykluczenie wpływu przypadku jako wyjaśnienie badanych rezultatów. Dlaczego zatem, w ogóle, używana jest statystyka? Czy badacz powinien wchodzić na drogę działania, gdy wynik badania, tak naprawdę, może być rezultatem przypadkowych zmian?" Jako uzupełnienie uwag Chow, Denis (2003) zauważa, że wielkość efektu jest tylko statystyką opisową, nieupoważniającą do wnioskowania. Co więcej, określa ona jedynie wielkość efektu w próbie i nie dostarcza żadnych informacji, jak wiarygodne jest to oszacowanie w populacji generalnej.

W szóstym wydaniu *Wytycznych* przygotowywania publikacji Amerykańskiego Towarzystwa Psychologicznego (APA, 2010) znalazłem dwa ciekawe sformułowania. Pierwsze to: „Assume that your reader has a professional knowledge of statistical methods" (załóż, że czytelnik twego artykułu ma profesjonalną wiedzę o metodach statystycznych). Czy w świetle przedstawionych w tej pracy wątpliwości występujących na fundamentalnym poziomie metodologii i filozofii statystyki matematycznej, jak również wątpliwości oraz kontrowersji także wśród matematyków, jest możliwe uzyskanie profesjonalnej wiedzy o metodach statystycznych wśród badaczy stosujących te metody? Założenie to wydaje mi się bardzo surrealistyczne. Drugie jest następujące: „For inferential statistical tests (e.g., t, F, and χ^2 tests), include the obtained magnitude or value of the test statistic, the degrees of freedom, the probability of obtaining a value as extreme as or more extreme than the one obtained (the exact p value), and the size and direction of the effect" (dla testów będących podstawą wnioskowania (np. testy t, F i χ^2) powinniśmy pokazać otrzymaną wielkość albo wartość statystyki testowej, liczbę stopni swobody, prawdopodobieństwo otrzymania wartości maksymalnej albo pewnego przedziału wyznaczonego przez wartość otrzymaną w teście, a także wielkość i kierunek efektu). Wymaganie podawania zarówno wartości statystyki będącej podstawą testu wraz z odpowiednią liczbą stopni swobody oraz prawdopodobieństwa odpowiadającego tej wartości statystyki jest bezsensowne i korzenie tego wymagania sięgają okresu sprzed przynajmniej 20 lat, gdy do oceny „istotności" wyniku wykorzystywane były tablice statystyczne. Służyły one do porównania obliczonej wartości statystyki z wartością krytyczną. W tej chwili każdy program statystyczny rachuje odpowiednie prawdopodobieństwo. Jak widać mity mają wyjątkowo długi żywot.

Na podstawie przejrzanego piśmiennictwa zaobserwowałem, niestety, jeszcze inną prawidłowość. W tekstach pisanych przez matematyków i statystyków nie pojawiają się metody oceny wielkości efektów, a najczęściej takie pojęcie w ogóle w tych pracach nie występuje. Natomiast w tekstach pisanych przez psychologów, socjologów, badaczy społecznych – oczywiście tam, gdzie jest sens wykorzystywać metody statystyczne – zawsze występuje wielkość efektu. Pocieszające są wyjątki w tym względzie, na przykład Chow (1996) czy Denis (2003).

BIBLIOGRAFIA

Agresti A. (1990). *Categorical Data Analysis*. New York: John Wiley and Sons.

Allen J., Le H. (2007). An additive measure of overall effect size for logistic regression models. *Journal of Educational and Behavioral Statistics*, 33, 416–441.

Anscombe F. J., Aumann R. J. (1963). A definition of subjective probability. *The Annals of Mathematical Statistics*, 34 (1), 199–205.

APA (2010). *Publication Manual*, 6th ed. Washington: American Psychological Association.

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

Blalock H. M. (1975). *Statystyka dla socjologów*. Warszawa: PWN.

Blume J. D. (2002). Likelihood methods for measuring statistical evidence. *Statistics in Medicine*, 21, 2563–2599.

Christensen R. (2005). Testing Fisher, Neyman, Pearson, and Bayes. *The American Statistician*, 59 (2), 121–126.

Chinn S. (2000). A simple method for converting an odds ratio to effect size for use in meta-analysis. *Statistics in Medicine*, 19 (22), 3127–3131.

Chow S. L. (1996). *Statistical Significance: Rationale, Validity and Utility*. London: Sage Publications.

Cohen J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Hillsdale: Lawrence Erlbaum Associates, Inc.

Cohen J. (1992). Statistical power analysis. *Current Directions in Psychological Sciences*, 1 (3), 98–101.

Denis D. J. (2003). Alternatives to null hypothesis significance testing. *Theory and Science*, 4 (1), 1–17.

Dienes Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspective on Psychological Science*, 6 (3), 274–290.

Dooling D. J., Danks J. H. (1975). Going beyond tests of significance: Is psychology ready? *Bulletin of the Psychonomic Society*, 5, 15–17.

Dudek B. (2007). Stres związany z pracą: teoretyczne i metodologiczne podstawy badań zależności między zdrowiem a stresem zawodowym. [W:] M. Górnik-Durose, B. Kożusznik (red.), *Perspektywy psychologii pracy* (s. 220–246). Katowice: Wydawnictwo Uniwersytetu Śląskiego.

Favreau O. E. (1997). Sex and gender comparison: Does null hypothesis testing create a false dichotomy? *Feminism and Psychology*, 7, 63–81.

Field A. (2009). *Discovering Statistics Using SPSS*, 3rd ed. London: Sage Publications.

Fisher R. A. (1935). The logic of inductive inference (with discussion). *Journal of the Royal Statistical Society*, 98 (1), 39–82.

Fisz M. (1969). *Rachunek prawdopodobieństwa i statystyka matematyczna*. Warszawa: PWN.

Greenland S., Maclure M., Schlesselman J. J., Poole C., Morgenstern H. (1991). Standardized regression coefficients: A further critique and review of some alternatives. *Epidemiology*, 2 (5), 387–392.

Greenland S., Schlesselman J. J., Criqui M. H. (1986). The fallacy of employing standardized regression coefficients and correlations as measures of effect. *American Journal of Epidemiology*, 123 (2), 203–208.

Greń J. (1968). *Modele i zadania statystyki matematycznej*. Warszawa: PWN.

Hilbe J. M. (2009). *Logistic Regression Models*. Boca Raton: Chapman and Hall/CRC.

Hoenig J. M., Heisey D. M. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. *The American Statistician*, 55 (1), 19–24.

Hosmer D. W., Lemeshow L. (1989). *Applied Logistic Regression*. New York: John Wiley and Sons.

Hubbard R., Armstrong J. S. (2006). Why we don't really know what "statistical significance" means: A major educational failure. *Journal of Marketing Education*, 28 (2), 114–120.

Hubbard R., Bayarri M. J. (2003). Confusion over measures of evidence (p's) versus errors (α's) in classical statistical testing. *The American Statistician*, 57 (3), 171–182.

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

Jeffreys H. (1961). *Theory of Probability*, London: Oxford University Press.

Jones L. V., Tukey J. W. (2000). A sensible formulation of the significance test. *Psychological Methods*, 5 (4), 411–414.

Karni E. (1993). A definition of subjective probabilities with state-dependent preferences. *Econometrica*, 61 (1), 187–198.

Kelley K., Preacher K. J. (2012). On effect size. *Psychological Methods*, 17 (2), 137–152.

Killeen P. R. (2005). An alternative to null-hypothesis significance tests. *Psychological Science*, 16 (5), 345–353.

Kline R. B. (2013). *Beyond Significance Testing. Statistics Reform in the Bahavioral Sciences*, 2nd ed. Washington: American Psychological Association.

Kołmogorow A. N. (1933). *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Berlin: Springer-Verlag. Za: H. Bauer (1968). *Probability Theory and Elements of Measure Theory*. New York: Holt, Rinehart and Winston, Inc.

Laplace P. S. (1812). *Theorie analytique des probabilites*. Paris: Courcier.

Lehmann E. L. (1993). The Fisher, Neyman-Pearson theories of testing hypotheses: One theory or two? *Journal of the American Statistical Association*, 88 (424), 1242–1249.

Lehmann E. L. (1995). Neyman's Statistical Philosophy. *Probability and Mathematical Statistics*, 15, 29–36.

Lenth R. V. (2007). Post hoc power: Tables and commentary. *Technical Report No. 378*, The University of Iowa, Department of Statistics and Actuarial Sciences, July, 1–13.

Levine T. R., Weber R., Hullett C., Park H. S., Lindsey L. L. M. (2008). A critical assessment of null hypothesis significance testing in quantitative communication research. *Human Communication Research*, 34, 171–187.

Lindgren B. W. (1962). *Statistical Theory*. New York: The Macmillan Co.

Lindquist E. F. ([1938] 1993). *A first course in statistics*. Cambridge: Houghton Miffilin. Za: C. J. Huberty. Historical origins of statistical testing practices: The treatment of Fisher versus Neyman-Pearson views in textbooks. *Journal of Experimental Education*, 61 (4), 317–333.

Machina M. J., Schmeidler D. (1992). A more robust definition of subjective probability. *Econometrica*, 60 (4), 745–780.

Magee L. (1990). R2 measures based on Wald and likelihood ratio joint significance tests. *The American Statistician*, 44 (3), 250–253.

Magiera R. (2007). *Modele i metody statystyki matematycznej*. Cz. II. *Wnioskowanie statystyczne*, wyd. 2 rozszerz. Wrocław: Oficyna Wydawnicza GiS.

Manthey J. (2010). *Elementary Statistics: A History of Controversy*. Boston: AMATYC 2010 Conference – Bridging Past to Future Mathematics, 11–14 November.

Menard S. (2000). Coefficients of determination for multiple logistic regression analysis. *The American Statistician*, 54 (1), 17–24.

Mises R. von (1936). *Wahrscheinlichkeit, Statistik und Wahrheit*. Wienna: Springer Verlag.

Nagelkerke N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, 78 (3), 691–692.

Neyman J. (1977). Frequentist probability and frequentist statistics. *Synthese*, 36, 97–131.

Neyman J., Pearson E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A*, 231, 289–337.

Za: E. L. Lehmann (1995). Neyman's statistical philosophy. *Probability and Mathematical Statistics*, 15, 29–36.

O'Keefe D. J. (2007). Post hoc power, observed power, a priori power, retrospective power, prospective power, achieved power: Sorting out appropriate uses of statistical power analyses. *Communications Methods and Measures*, 1 (4), 291–299.

Onwuegbuzie A. J., Leech N. L. (2004). Post hoc power: A Concept whose time has come. *Understanding Statistics*, 3 (4), 201–230.

Papoulis A. (1972). *Prawdopodobieństwo, zmienne losowe i procesy stochastyczne*. Warszawa: Wydawnictwa Naukowo-Techniczne.

Rao C. R. (1982). *Modele liniowe statystyki matematycznej*. Warszawa: PWN.

Rasch D. (2012). Hypothesis testing and the error of the third kind. *Psychological Test and Assessment Modeling*, 54 (1), 90–99.

Roberts S., Pashler H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, 107 (2), 358–367.

Rodgers J. L. (2010). The epistemology of mathematical and statistical modeling. A quiet methodological revolution. *American Psychologist*, 65 (1), 1–12.

Rosenthal R. (1991). *Metaanalytic Procedures for Social Research*, 2nd ed. Newbury Park: Sage.

Rosnow R. L., Rosenthal R. (2005). *Beginning behavioural research: A conceptual primer*, 5th ed. Englewood Cliffs NJ: Pearson/Prentice Hall.

Royall R. (2000). On the probability of observing misleading statistical evidence (with comments). *Journal of the American Statistical Association*, 95 (451), 760–780.

Royall R. (1997). *Statistical Evidence. A Likelihood Paradigm*. London: Chapman and Hall/CRC.

Sedlmeier P., Gigerenzer G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, 105 (2), 309–316.

Seltman H. J. (2014). *Experimental design and analysis. Chapter 12: Statistical power*, http://www.stat.cmu.edu/~hseltman/309/Book/Book.pdf [dostęp: 10.12.2014].

Silvey S. D. (1978). *Wnioskowanie statystyczne*. Warszawa: PWN.

Sink C. A., Mvududu N. H. (2010). Statistical power, sampling, and effect sizes: Three keys to research relevancy. *Counseling Outcome Research and Evaluation*, 1 (2), 1–18.

Sterne J. A. C. (2002). Teaching hypothesis tests – time for significant change? *Statistics in Medicine*, 21 (7), 985–994.

Szymczak W. (2010). *Podstawy statystyki dla psychologów*. wyd. 2 popr. Warszawa: Difin.

Tabachnick B. G., Fidell L. S. (2007). *Using Multivariate Statistics*, 5th ed. Boston: Pearson Education, Inc.

Thalheimer W., Cook S. (2002). How to calculate effect sizes from published research articles: A simplified methodology, http://work-learning.com/effect_sizes.htm [dostęp: 28.08.2012].

Thompson B. (1994). The concept of statistical significance testing. *Practical Assessment, Research and Evaluation*, 4, 5.

Valentine J. C., Cooper H. (2003). *Effect Size Substantive Interpretation Guidelines: Issues in the Interpretation of Effect Sizes*. Washington: What Works Clearinghouse.

Volker M. A. (2006). Reporting effect size estimates in school psychology research. *Psychology in the Schools*, 43 (6), 653–672.

Williams R. H., Zimmerman D. W. (1989). Statistical power analysis and reliability of measurement. *Journal of General Psychology*, 116 (4), 359–369.

Zubrzycki S. (1970). *Wykłady z rachunku prawdopodobieństwa i statystyki matematycznej*. Warszawa: PWN.

Wiesław Szymczak

THE CONCEPT OF SIZE EFFECT IN THE LIGHT OF NEYMAN-PEARSON'S THEORY OF TESTING STATISTICAL HYPOTHESIS

Abstract. The aim of this study is to draw the attention of researchers using statistical methods in the analysis of the results of their research on the combination of two different theories testing statistical hypothesis, Fisher's theory and Neyman-Pearson's theory. Including in the presently used statistical instruments, ideas of both of these theories, causes that the vast majority of researchers without a moment's thought, acknowledge that the smaller the probability the stronger relationship. The study presents the weaknesses of Neyman-Pearson's theory and the resulting problems with decision-making as a result of the conducted tests. These problems have become a justified quest for less unreliable solutions, however, the proposed measures of the size effect as using on one hand dogma about the relationship between the degree of probability in the test and the strength of dependence, on the other, lack of any theoretical basis of this solution, seem to be another pseudo solution to actual problems. Moreover, the use of measures of size effect seems to be an attempt to free researchers from the profound thinking about the results obtained from the statistical analysis. A trivial recipe was established: the corresponding value of the measures instantly implies the strength of the relationship – this approach seems unworthy of the researcher.

Keywords: theories of statistical hypothesis testing, probability, power of test, empirical power of test, effect size.