PRZEGLĄD PSYCHOLOGICZNY, 2000, TOM 43, NR 1, 29-38

Nowe ujęcie problematyki mocy testów statystycznych

Elżbieta Aranowska, Jolanta Rytel[1]
Wydział Psychologii Uniwersytetu Warszawskiego

NEW APPROACH TO THE PROBLEMS OF STATISTICAL TESTS POWER

Summary. This article presents a discussion of the statistical tests power concept based the Neyman-Pearson approach to the null hypothesis testing. The notion of power is presented from the classical point of view. Necessity to include two inference errors in planning research is postulated and the examples of calculating the test power as well as the sample size needed to secure the desired power are described. Moreover a revised concept of power based on Kaiser's (1960) proposal of 3-alternative hypothesis testing is discussed and suggestions are offered for assuming of Type I and Type II errors values.

KLASYCZNA DEFINICJA MOCY TESTU I GŁÓWNE WŁASNOŚCI MOCY

W paradygmacie Neymana-Pearsona rozważa się dwa rodzaje błędów związanych z procedurą wnioskowania statystycznego. W centrum zainteresowania badaczy przede wszystkim leży błąd pierwszego rodzaju, wiążący się z podjęciem błędnej decyzji polegającej na odrzuceniu prawdziwej hipotezy zerowej. Rzadko kiedy badacze pamiętają o tym, że błąd drugiego rodzaju jest immanentnie wpisany w schemat wnioskowania statystycznego. Z jednej strony „w większości problemów stosunek do obu hipotez – zerowej i alternatywnej – jest różny. Zwykle pyta się o to, czy istnieją dostateczne dowody, aby można było odrzucić hipotezę zerową" (Aranowska, Rytel, 1997, s. 250). Nieodrzucenie hipotezy zerowej nie wpływa na zmianę sposobu jej traktowania (w dalszym ciągu badacz jest jednakowo niepewny co do jej statusu). Natomiast odrzucenie hipotezy daje badaczowi możliwość oceny prawdopodobieństwa jej fałszywości na wysokim poziomie, równym $1 - \alpha$. „A zatem błąd I rodzaju odgrywa większą rolę niż błąd II rodzaju. Jego popełnienie jest bardziej „kosztowne" (Aranowska, Rytel, 1997, s. 250). Z drugiej strony w praktyce badawczej w większości przypadków poprzestaje się na fragmentarycznym, anachronicznym odwołaniu się do teorii wnioskowania, to znaczy na operowaniu wyłącznie błędem pierwszego rodzaju (por. Sedlmeier, Gigerenzer, 1989). Byłoby to właściwe, gdyby badacz stawiał i poddawał weryfikacji jedną tylko hipotezę, a mianowicie hipotezę zerową, tak jak ma to miejsce w Fisherowskim paradygmacie wnioskowania statystycznego. Stawiając jednak dwie hipotezy, zerową i alternatywną, badacz zmuszony jest do rozważania, a tym samym do uwzględniania obu rodzajów błędów: zarówno błędu pierwszego rodzaju, jak i błędu drugiego rodzaju, rozumianego jako błąd decyzyjny badacza polegający na przyjęciu fałszywej hipotezy zerowej. W tabeli zilustrowano możliwe decyzje badacza.

Błędy decyzyjne, prawdopodobieństwa ich popełnienia
w trakcie wnioskowania statystycznego oraz prawdopodobieństwa decyzji

Decyzja badacza $D = D_1\ 4\ D_2$	Hipoteza zerowa prawdziwa H_0^P	Hipoteza zerowa fałszywa H_0^F
Przyjąć H_0 D_1	Decyzja właściwa, dla której błąd wynosi 0 $p(D_1 \mid H_0^P) = 1 - \alpha$	Błąd II rodzaju wielkości ß $p(D_1 \mid H_0^F) = 1 - ß$
Odrzucić H_0 D_2	Błąd I rodzaju wielkości α $p(D_2 \mid H_0^P) = 1 - \alpha$	Decyzja właściwa, dla której błąd wynosi 0 $p(D_2 \mid H_0^F) = 1 - ß$
	$p(D) = 1$	$p(D) = 1$

Czytelnikowi może wydać się dziwne, iż w komórkach tabeli ilustrujących właściwe decyzje badacza prawdopodobieństwo związane z tymi decyzjami nie jest równe jedności, choć błędy obciążające te decyzje mają wartość zero. Dzieje się tak, dlatego że status logiczny hipotezy zerowej (prawda *vs.* fałsz) nigdy nie jest znany.

1 Adres do korespondencji: Wydział Psychologii, Uniwersytet Warszawski, ul. Stawki 5/7, 00-183 Warszawa; e-mail: aranella@sci.psych.uw.edu.pl.

Gdyby było wiadomo, że hipoteza zerowa jest prawdziwa bądź fałszywa, to przyjęcie jej w pierwszej sytuacji lub też odrzucenie – w drugiej, nie byłoby obciążone żadnym błędem, zaś obciążone błędami byłyby decyzje polegające na odrzuceniu prawdziwej (wartość błędu $\alpha = 1$) lub też przyjęciu fałszywej (wartość błędu $\beta = 1$) hipotezy zerowej. Jednakże wnioskowanie indukcyjne prowadzi do decyzji podejmowanych w warunkach niepewności co do statusu hipotezy zerowej. Decyzje podejmowane są jedynie na podstawie rezultatów uzyskanych z badania *n*-elementowej próby losowej za pomocą wybranego narzędzia statystycznego. Zatem zdolność tego narzędzia do wykrywania rzeczywistych relacji pomiędzy analizowanymi zmiennymi (np. różnicy między wartościami parametrów) nigdy nie jest absolutna, jako że obserwowane na poziomie próby następstwa analizowanych praw przyrody są fragmentaryczne i jako takie nie zawsze „prawdziwie" odzwierciedlają zachodzące w przyrodzie relacje. W tych warunkach żadna z podejmowanych przez badacza decyzji nie może być decyzją całkowicie pewną.

Podejmując decyzję w warunkach niepewności, badacz zmuszony jest do brania pod uwagę równocześnie obydwu możliwości: prawdziwości lub fałszywości hipotezy zerowej. Zatem poziom zaufania badacza co do trafności podejmowanych decyzji wyrażać mogą jedynie prawdopodobieństwa warunkowe (por. tabela). Wartości tych prawdopodobieństw określają pewność badacza zmniejszoną o odpowiedni (zrelatywizowany do rozważanego warunku) błąd wnioskowania.

Definicja mocy testu statystycznego

Moc testu statystycznego rozumiana jest jako prawdopodobieństwo nieodrzucenia hipotezy alternatywnej pod warunkiem, że jest ona prawdziwa. Innymi słowy, jest ona prawdopodobieństwem odrzucenia hipotezy zerowej, przy założeniu, że hipoteza ta jest fałszywa. Z analizy tab. 1 wynika, że moc zdefiniowana jest jako wielkość równa $1 - \beta$ (por. decyzja D_2). Zatem maksymalizując moc testu statystycznego minimalizuje się zarazem prawdopodobieństwo popełnienia błędu decyzyjnego drugiego rodzaju, czyli β. Interpretując moc testu jako zdolność testu statystycznego do odrzucania hipotez fałszywych dochodzimy do wniosku, że im wyższa moc, tym mniejsze ryzyko nieodrzucenia fałszywej hipotezy zerowej.

W naukach społecznych hipotezy operacyjne odnoszą się z reguły do kierunku wpływu jednej zmiennej na drugą, a nie do rozmiaru tego wpływu, inaczej realnego efektu wpływu wyróżnionej zmiennej na inną. Tak więc badacz na poziomie stawianych przez niego problemów *implicite* oczekuje odpowiedzi, że np. „coś będzie większe od czegoś" (średnia poziomu wykonania jednego zadania od innego) czy też: „coś będzie lepsze od czegoś" itp. (Leventhal, Huynh, 1996). Wielokrotnie zdarza się, że badacz nie wykorzystuje dostępnej mu wiedzy (teoretycznej bądź praktycznej), dotyczącej wielkości różnic (np. średnich) świadczących o rzeczywistym efekcie relacji między wskaźnikami konkretnych konstruktów. A przecież uzyskane w badaniu różnice, choć istotne statystycznie, ale jednocześnie mniejsze od realnie (merytorycznie) traktowanych jako dostateczne, nie powinny stanowić podstawy do refleksji odnośnie do wartości efektu eksperymentalnego czy − ogólniej – badawczego. Nie dość, że *a priori* najczęściej nie określa się wielkości spodziewanego efektu, to z reguły *a posteriori* nie stosuje się znanych miar pozwalających na oszacowanie wielkości efektu ujawnionego w badaniu (por. Kirk, 1996).

Moc testu jako prawdopodobieństwo nieodrzucenia prawdziwej hipotezy alternatywnej może być interpretowane jako miara wykrycia realnego efektu wpływu analizowanych zmiennych na inne zmienne, w szczególności można je uznać za miarę zmiany wartości zmiennej zależnej spowodowaną oddziaływaniem zmiennej niezależnej.

Biorąc pod uwagę współzależność zachodzącą pomiędzy mocą testu a wielkością błędu drugiego rodzaju, można sformułować jej kilka własności.

1. Moc zmienia się odwrotnie względem zmian wartości błędu II rodzaju (β).
2. Moc zmienia się w tym samym kierunku, co wielkość próby; zwiększając liczebność próby, zwiększa się moc testu.
3. Moc zmienia się w tym samym kierunku, co wielkość realnego efektu zmiennej. W przypadku dwóch badań przeprowadzonych na próbach o tej samej liczebności, przy tym samym poziomie istotności α, moc testu jest wyższa dla dużego efektu w porównaniu z małym. Innymi słowy, „łatwiej" można ujawnić w badaniach większe efekty.
4. Moc zmienia się w tym samym kierunku, co wielkość błędu I rodzaju (α). Im mniejszy poziom istotności, tym mniejsza moc testu.

Powyższe zależności implikują, że każda z czterech zmiennych: moc testu, wielkość próby, wielkość efektu i poziom istotności może być wyznaczona na podstawie trzech pozostałych. Najczęściej wyznaczaną wielkością – przy założonej liczebności próby – jest moc testu. I odwrotnie, przy założonym, dostatecznie wysokim poziomie mocy testu wyznacza się niezbędną wielkość próby.

W literaturze statystycznej rekomendowanym poziomem mocy testu jest wartość większa lub równa 0,8 (Cohen, 1965). Oznacza to, że przy ustalonym poziomie istotności α wielkość błędu II rodzaju nie może być większa niż 0,2.

NOWE UJĘCIE PROBLEMATYKI MOCY TESTÓW STATYSTYCZNYCH

Wyznaczanie mocy testu ważne jest na dwóch etapach realizacji procesu badawczego: a) w trakcie planowania badania; b) w trakcie interpretowania uzyskanych wyników analizy danych.
W pierwszej sytuacji z reguły poszukuje się liczebności próby wystarczającej do wykrycia efektu o zakładanej przez badacza wielkości przy ustalonej, wysokiej mocy wykorzystywanego testu statystycznego. Natomiast w trakcie interpretacji wyników badacz ma możliwość skonfrontowania efektu założonego z otrzymanym. Może się wtedy okazać, iż obydwie wielkości znacznie od siebie odbiegają. Jeśli np. efekt otrzymany jest znacznie mniejszy od zakładanego i nieistotny statystycznie, ważne jest, aby badacz zrewidował rzeczywistą moc zastosowanego testu. Analizowanie mocy w obydwu powyższych sytuacjach jest szczególnie ważne, gdy podejmowane badania mają charakter eksploracyjny i badacz z reguły nie może sformułować precyzyjnych przypuszczeń co do natury zależności między analizowanymi zmiennymi.

SZACOWANIE MOCY ZGODNIE Z JEJ KLASYCZNYM UJĘCIEM
NA ETAPIE PLANOWANIA BADANIA

Przykład wyznaczania mocy testu dla próby o określonej liczebności
Zgodnie z powyższym, na etapie planowania badania jednym z najważniejszych pytań jest pytanie o niezbędną wielkość próby wystarczającą do uzyskania zadowalającej mocy testu. Rozwiązanie tego problemu zostało zilustrowane przykładem, w którym statystyką wykorzystywaną do weryfikacji hipotezy zerowej postaci $\mu = \mu_0$ jest statystyka *z*. Załóżmy, że badacz chce ocenić skuteczność nowej metody treningu pamięci. Przeciętny czas opanowania materiału pamięciowego przy posługiwaniu się dotychczasową metodą treningu wynosił 50 sekund, $\mu_0 = 50$, natomiast odchylenie standardowe było równe 16 sekund. Przy założeniu, że zmienna „czas zapamiętywania" ma rozkład normalny oraz znanych wartościach statystyk w próbie po zastosowaniu nowej metody treningu, badacz ma możliwość oceny jej efektywności. Opierając się na znajomości specyfiki procesów pamięciowych i wiedzy dotyczącej wykorzystywanego materiału (specyfiki narzędzia psychologicznego) badacz zakłada, że efektywność nowej metody treningu zostanie wykazana, gdy przeciętny czas konieczny do opanowania materiału zmniejszy się co najmniej o 3 sekundy. Toteż do weryfikacji hipotezy o braku zmian czasów przeciętnych obydwu metod treningu badacz posłuży się testem jednostronnym, przy ustalonym poziomie istotności, np. $\alpha = 0,05$.
Na tym etapie planowania badania badacz staje przed problemem określenia liczby elementów próby wystarczającej do ujawnienia założonego efektu o wielkości co najmniej 3 sekund. Innymi słowy, badacz rozważa, czy próba na przykład 25-elementowa jest próbą o wystarczającej wielkości dla uzyskania zadowalającej mocy testu. Gdyby badacz zastosował test *z* postaci:
(1)
szacowanie mocy przebiegałoby w sposób zgodny ze standardową procedurą wyznaczania wielkości błędu II rodzaju.
1. Należy wyznaczyć graniczną wartość w rozkładzie z próby średnich czasów zapamiętywania przy założeniu prawdziwości hipotezy$_{\text{zerowej}}$ (por. rys.1):
(2)
(3) gdzie
(4)
(5)
Stąd
2. Uzyskaną wartość graniczną ocenić w kategoriach liczby odchyleń standardowych od średniej $\mu = \mu_0 - 3 = 50 - 3 = 47$, to znaczy wystandaryzować ją:
(6)
Stąd
Moc testu w tym przypadku stanowi wartość dystrybuanty dla otrzymanego wyniku wystandaryzowanego.

(7) moc = $P(_{\text{otrzym.}} < 44, 736) = P(z_{\text{otrzym.}} < -0, 7075) = \Phi(-0,71)$.
Moc = 0,2389.
W konsekwencji: $\beta = 1 - \text{moc} = 0,7611$.

P

Rys. 1. Ilustracja mocy testu (szere pole) dla próby o rozmiarze $n = 25$

Powyższe wyniki dobitnie dokumentują fakt nadmiernego optymizmu badacza co do wstępnie zakładanej, zdecydowanie zbyt małej liczebności próby. Badanie w warunkach tak ogromnej wartości błędu II rodzaju, a tym samym tak dramatycznie niskiej wartości mocy testu, nie może zostać przeprowadzone.
Warto zastanowić się, czy zwiększenie liczebności próby do np. 100 elementów, co uznawane jest w większości podręczników statystyki dla badaczy w naukach społecznych za całkowicie wystarczający rozmiar, w powyższym przypadku pozwoli na zadowalającą poprawę mocy testu. Analogicznie do powyższego można wyznaczyć wartości kolejnych statystyk:
błąd standardowy średniej
wartość graniczną w rozkładzie z próby średnich:

wystandaryzowaną wartość graniczną:

$\text{moc} = P(z_{\text{otrzym.}} < 0{,}23) = \Phi(0{,}23) = 0{,}591;$
$ß = 1 - \text{moc} = 0{,}4090.$

Widać, że mimo czterokrotnego zwiększenia liczebności próby, czyli dwukrotnego spadku wartości błędu standardowego średniej, nadal nie udało się otrzymać zadowalającej wartości mocy testu, a tym samym możliwej do zaakceptowania wartości błędu II rodzaju.

Przykład wyznaczania liczebności próby przy ustalonej wartości mocy testu
Niewątpliwie problem wymaga przeformułowania i zapytania wprost: jakiego rozmiaru powinna być próba, aby przewidywany co najmniej 3-sekundowy efekt spadku przeciętnego czasu zapamiętywania został wykryty z dużym prawdopodobieństwem, to znaczy przy mocy rzędu 0,8 (por. rys. 2). Łatwo pokazać, że prawdziwy jest wzór:
(8) $n = [\sigma^2 \times (z_{\text{kryt.}} - z_{\text{otrzym.}})^2] / (\mu_0 - \mu)^2;$
(9) $z_{\text{otrzym.}} = z_{0{,}80} = 0{,}84.$
Stąd:
$n = [16^2 \times (-1{,}645 - 0{,}84)^2] / (47 - 50)^2] = 177.$

P

Rys. 2. Ilustracja mocy testu (szare pole) wielkości równej 0,8

Dopiero 7-krotne zwiększenie liczebności próby względem pierwotnego zamysłu badacza gwarantuje wiarygodne wnioskowanie statystyczne. Badacz, planując badanie, posłużył się założeniami odnośnie do obydwu rodzajów błędów wnioskowania zgodnie z paradygmatem Neymana-Pearsona. Były to jednak założenia liberalne, tzn. wartości obydwu błędów były wartościami najwyższymi z możliwych (zgodnie z przyjętą, dopuszczalną

w naukach empirycznych granicą). Warto podkreślić, że jakiekolwiek bardziej rygorystyczne założenia odnośnie do wielkości błędów (zmniejszenie wartości dowolnego z nich lub obydwu równocześnie) prowadziłyby do znacznie większej wartości *n*, do większego rozmiaru losowanej próby. W podobny jak przedstawiony wyżej sposób, posługując się rozkładami dowolnych statystyk z prób, można wyznaczać liczebności prób wystarczające dla uzyskania zadowalających wartości mocy testów. Jeżeli badacz rozważa możliwość wykorzystania do analizy danych wielu metod statystycznych, spośród wielkości prób wyznaczonych dla każdej z nich, powinien wybrać próbę o rozmiarze maksymalnym.
Komentarza wymaga – w kontekście wyżej opisanej procedury obliczeniowej – wprowadzone wcześniej założenie o normalnym rozkładzie prawdopodobieństwa analizowanej zmiennej „czas zapamiętywania". Naturalnie wyłącznie przy takim założeniu wyznaczanie rozmiaru próby w opisany sposób ma sens. Jednakże wielokrotnie przed zrealizowaniem badania nie wiadomo, czy założenie to ma zastosowanie w danym przypadku. Chociaż teoria odporności testów statystycznych (por. np. Zieliński, 1993) dopuszcza pogwałcenie tego założenia w testach weryfikujących hipotezy o równości średnich z reguły do momentu zachowania kształtu krzywej, to w sytuacji, gdy badacz nie wie nic na temat typu rozkładu zmiennej zależnej w interesującej go populacji, jedyną dyrektywą wyznaczania liczebności próby powinno być odwołanie się do nierówności Czebyszewa (por. Aranowska, Rytel, 1997). Jeśli badacz potrafi w przybliżeniu ocenić wielkość efektu (spadku lub wzrostu średnich) oraz rozrzut wyników (wielkość wariancji zmiennej), może rozważać możliwość wykorzystania wzoru (8) do określenia rozmiaru próby, traktując go jednak nieufnie, tzn. po przeprowadzeniu badania winien bezwzględnie po raz wtóry skontrolować moc, z jaką kolejne testy statystyczne były stosowane (po skonfrontowaniu obydwu efektów badania: założonego i otrzymanego; por. pkt 1).

ZREWIDOWANA DEFINICJA MOCY TESTU STATYSTYCZNEGO

Implikacje kierunkowego stawiania hipotez alternatywnych

Problemem szczególnej wagi, którego nie sposób pominąć w tym miejscu, jest możliwość wyciągania wniosków z wyników badań ujawnionych w trakcie testowania hipotezy zerowej testem jednostronnym. Wymaga ona kilku komentarzy.
Pierwszy dotyczy powszechnie akceptowanej sugestii, zawartej w literaturze metodologicznej, odnoszącej się do niestawiania hipotez kierunkowych, gdy badacz eksploruje problem. A przecież właśnie z metodologicznego punktu widzenia jest to sugestia nietrafna. Jak wiadomo, to dane z badań w ostatecznym rozrachunku potwierdzają lub obalają stawiane hipotezy, gdyż to właśnie dane (wyniki próby) są następstwem ewentualnie zachodzących praw przyrody, a zatem tylko z danych można czerpać wiedzę o charakterze praw i kierunku relacji. Dodatkowo ze statystycznego punktu widzenia stawianie alternatywnych hipotez kierunkowych – przy tej samej wielkości błędu wnioskowania I rodzaju – umożliwia minimalizację wielkości błędu II rodzaju, a zatem zwiększa się moc testu statystycznego. Hipotezy zerowe z reguły powinny być „obalane" na rzecz kierunkowych hipotez alternatywnych.
Drugie zagadnienie wiąże się z wiarygodnością samych danych. Jest to problem tak ważki, że został podjęty na nowo po prawie dwudziestu latach milczenia od chwili ukazania się znamiennej pracy Kaisera (1960) na temat kierunkowych decyzji statystycznych. W przywołanej pracy Kaiser wprowadza pojęcie kierunkowego testu dwustronnego, dopuszczając równocześnie możliwość analizowania obydwu kierunków rozkładu statystyki testu (przy poziomie istotności równym $\alpha/2$) oraz możliwość wypowiadania wniosków „kierunkowych". Po raz pierwszy zatem zostało postawione pytanie o przyjęcie poprawnego kierunku wnioskowania (uznanie za prawdopodobny kierunku przedstawionego w jednej z hipotez alternatywnych).
W konsekwencji procedurę wnioskowania statystycznego rozszerza się o analizę trzeciego rodzaju błędu, błędu polegającego na odrzuceniu hipotezy zerowej w złym kierunku, o wielkości γ (Kaiser, 1960; Leventhal, Huynh, 1996).

Błąd III rodzaju i zrewidowana definicja mocy testu

Gdy zaakceptuje się logikę testowania hipotezy zerowej zaproponowaną przez Kaisera, wielkość tego błędu – γ – nie może być większa niż $\alpha/2$.
Wprowadzenie do wnioskowania błędu III rodzaju naturalnie nie pozostaje bez znaczenia dla mocy testu. Moc testu zdefiniowana zostaje jako:
(10) $\text{moc} = 1 - \beta - \gamma$.

Analizując wzór (10) dochodzimy do wniosku, że w nowym ujęciu moc nie zmienia się w sposób znaczny, skoro $\max \gamma = \alpha/2$.

Relacje między trzema rodzajami błędów wnioskowania

Ostatnim problemem wymagającym komentarza jest związek wszystkich rodzajów błędów, a dokładniej wpływ błędu pierwszego rodzaju na moc testu. Z faktu, że im wyższa wartość błędu pierwszego rodzaju, tym niższa wartość błędu drugiego rodzaju, nie wynika jeszcze, jak szybko ß maleje. Zależy to od wielu czynników, m.in. od liczebności próby i wielkości badanego efektu. Ponieważ γ nie przekracza α/2, nie wiadomo (zważywszy na (10)), do jakiego stopnia ma sens zwiększanie poziomu istotności wnioskowania przy minimalizacji ß.

Sensowne wydaje się wyznaczenie maksymalnej wielkości α w modelu trójdecyzyjnym, przyjmując za Cohenem (1965) minimalną wartość mocy testu równą 0,8 i maksymalną wartość γ.

Toteż na podstawie (10):

(11) $1 - \text{moc} = \alpha/2 + ß$.

W szczególności, gdy moc = 0,8,

$\alpha/2 + ß = 0{,}2$.

Natomiast, gdy moc = 0,9,

$\alpha/2 + ß = 0{,}1$.

Zatem maksymalne dopuszczalne α, gdy moc jest najmniejsza z przedziału wartości dopuszczalnych, nie może przekraczać:

$\max \alpha = 0{,}4 - 2ß$.

Odwrotnie wielkość błędu II rodzaju z poziomem istotności wiąże równanie:

$ß = 0{,}2 - \alpha/2$.

Jeżeli przy minimalnej akceptowalnej mocy testu równej 0,8, gdy dopuszcza się wartość ß nie większą niż 0,1, to maksymalne $\alpha = 0{,}2$. Przy założonej minimalnej mocy testu równej 0,8 błąd ß będzie większy od α wyłącznie dla α mniejszego od 0,13.

Przedstawione wyżej rozważania odbijają się oczywiście na wielkości próby niezbędnej do potwierdzenia założonego efektu. Wielkość poziomu istotności wzbudza w literaturze statystycznej wiele emocji (por. Cohen, 1994). Dopiero jednak trójdecyzyjny model wnioskowania umożliwia wskazanie *explicite* granic wartości dwu najważniejszych błędów wnioskowania I i II rodzaju.

Jeżeli błąd III rodzaju ma wartość mniejszą od możliwej górnej granicy równej α/2, warunek $ß \geq \alpha$ będzie spełniony dla coraz większych względem 0,13 wartości α (np. gdy $\gamma = \alpha/4$, poziom istotności może być równy nawet 0,16). Zwiększając moc testu, warunek $ß \geq \alpha$ spełniony zostanie dla wartości α znacznie mniejszych od wartości granicznej równej 0,13. Dla mocy testu równej 0,9, ß będzie większe od α dla wartości $\alpha \leq 0{,}065$, czyli dla wartości granicznej, tutaj dwukrotnie mniejszej niż w przypadku mocy równej 0,8. Zatem nie mityczny poziom istotności równy 0,05, lecz przy wnioskowaniu rygorystycznym $\alpha = 0{,}065$ stanowi sensowną górną granicę wartości dopuszczalnych, gdy badacz przyjmuje rozszerzony o błąd III rodzaju paradygmat Neymana-Pearsona. Parafrazując stwierdzenie Rosnowa i Rosenthala (1989) o tym, iż Bóg równie mocno kocha $\alpha = 0{,}06$, jak i $\alpha = 0{,}05$, w świetle powyższych rozważań należy sądzić, że najmocniej kocha jednak $\alpha = 0{,}065$.

BIBLIOGRAFIA

Aranowska, E., Rytel, J. (1997). Istotność statystyczna – co to naprawdę znaczy? *Przegląd Psychologiczny*, 40, 249-260.

Cohen, J. (1965). Some statistical issues in psychological research. [W:] B. B. Wolman (red.), *Handbook of clinical psychology*. New York: Academic Press, 92-121.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, 49, 997-1003.

Kaiser, H. (1960). Directional statistical decisions. *Psychological Review*, 67, 160-167.

Kirk, R. (1996). Practical significance: a concept whose time has come. *Educational and Psychological Measurement*, 56, 746-759.

Leventhal, L., Huynh, C. (1996). Directional decisions for two-tailed tests: power, error rates, and sample size. *Psychological Methods*, 1, 278-292.

Rosnow, R. Rosenthal R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist*, 44, 1276-1284.

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

Zieliński, R. (1993). Koncepcja odporności w statystyce. *Psychologia Matematyczna*, 5, 7-14.