PSYCHOLOGICZNE ZESZYTY NAUKOWE
PÓŁROCZNIK INSTYTUTU PSYCHOLOGII UNIWERSYTETU ZIELONOGÓRSKIEGO
NR 1/2017

Człowiek w całokształcie życia: rodzina – edukacja – praca

ŁUKASZ BUDZICZ
Uniwersytet Zielonogórski

INTERPRETACJA STATYSTYK W ARTYKUŁACH NAUKOWYCH – WSKAZÓWKI DLA PRAKTYKÓW

WSTĘP

Zawód psychologa jest zawodem zaufania publicznego. Wynika z tego szereg konsekwencji. Jedną z nich jest konieczność aktualizowania przez psychologa swojej wiedzy, tak aby w swojej praktyce zawodowej mógł wykorzystywać metody, techniki i narzędzia najbardziej zgodne ze współczesnym stanem wiedzy naukowej. W szczególności powinien być na tyle kompetentnym „konsumentem" wiedzy naukowej, żeby nie przyswajać, nierzadkich w tej dziedzinie, „cudownych", ale w istocie pseudonaukowych sposobów działania.

Skąd jednak psycholog ma wiedzieć, jaki jest „współczesny stan wiedzy naukowej"? Podstawowym źródłem wiedzy naukowej w psychologii są artykuły ukazujące się w recenzowanych czasopismach naukowych. Warto sięgać też po monografie naukowe (w języku potocznym nazywane „książkami"), szczególnie te, które są wydane przez czołowe światowe wydawnictwa akademickie. Jednak monografie, trzeba to podkreślić wyraźnie, nie są na ogół źródłem podstawowej wiedzy, lecz pełnią funkcje przeglądowe i podsumowujące. Podstawowym źródłem wiedzy są prawie zawsze artykuły w czasopismach naukowych.

Największą przeszkodą, na którą może trafić psycholog-praktyk przy czytaniu artykułów empirycznych, są pojawiające się w nich techniczne elementy analizy statystycznej. Na studiach psychologicznych kurs z metodologii i statystyki jest co prawda obowiązkowy, ale niestety, niektóre kursy nie podejmują zagadnień interpretacji danych

statystycznych, a jedynie uczą mechanicznego przeprowadzania obliczeń. A nawet jeśli te kursy uczą sensownego interpretowania, to często potem na samych studiach nie analizuje się artykułów empirycznych, a nieużywana wiedza zanika.

Wychodząc naprzeciw potrzebom praktyków chcących czytać artykuły naukowe, tekst ten ma być w zamyśle względnie krótkim przewodnikiem wprowadzającym do tematyki interpretowania danych statystycznych. Z mojego osobistego doświadczenia wynika, że nie trzeba być zawodowym statystykiem, żeby być w stanie dokonywać sensownych interpretacji, szczególnie takich, które są użyteczne na potrzeby praktyki. Podkreślę, że artykuł ten nie jest wyczerpującym kursem statystyki. Zachęcam Czytelnika do rozwijania wiedzy, na rynku są dostępne liczne wartościowe książki, szczególnie polecam podręczniki B. Kinga i E. Miniuma (2009) oraz P. Francuza i R. Mackiewicza (2007).

Zanim przejdę jednak do właściwej części, wspomnę o jeszcze innej trudności związanej ze zdobywaniem podstawowych źródeł wiedzy i sposobach jej przezwyciężenia. Wiele tekstów naukowych nie jest ogólnodostępnych, a mówiąc ściślej, są dostępne po poniesieniu wysokich kosztów (rzędu nawet 40 dolarów za jeden artykuł). Niestety, choć badania naukowe są zwykle finansowane ze środków publicznych, to czasopisma, w których są przedstawiane, często są wydawane przez wielkie prywatne (a więc komercyjne) konglomeraty wydawnicze (np. Elsevier, Springer, Wiley). Nie oznacza to jednak, że bezpłatny dostęp do takich artykułów jest zupełnie niemożliwy. Po pierwsze wielu badaczy umieszcza pełne wersje swoich tekstów na własnej stronie internetowej lub na wyspecjalizowanych serwisach (szczególnie researchgate.net i academia.edu). Według mojego rozeznania około 40-50% nowych artykułów naukowych można pozyskać tą drogą. Po drugie osoby, które mieszkają w dużych ośrodkach, mogą zapisać się do bibliotek uniwersyteckich, uiszczając stosukowo niewysokie opłaty, i ściągać lub drukować artykuły, korzystając z komputerów w czytelniach. Wreszcie jest jeszcze trzecia możliwość, mianowicie można napisać wiadomość e-mailową do autora artykułu z prośbą o przesłanie cyfrowej wersji tekstu. Bardzo często autorzy odsyłają artykuły, o które zostali poproszeni. Nadmienię jeszcze, że w ostatnich latach coraz większego znaczenia nabierają czasopisma w formule tak zwanych *open access*. Artykuły z tych czasopism są ogólnodostępne dla każdego, kto ma dostęp do Internetu. Można się spodziewać, że rola tych czasopism będzie rosła.

W ostatecznym rozrachunku, jeśli wszystkie opisane sposoby dotarcia do interesującego nas artykułu zawiodą, zawsze ogólnodostępne są tytuły i abstrakty, które pozwolą nam zorientować się w ogólnych wnioskach z badania.

Oczywiście, to nie są jedyne przeszkody. Inną jest bariera językowa. Olbrzymia większość wartościowych w obiegu naukowym tekstów ukazuje się w języku angielskim.

Trudno jednak przyjąć argument, że ze względu na barierę językową można nie czytać. Znajomość języka angielskiego jest konieczna w różnych zawodach wymagających wysokich kompetencji. Dodatkowo w tekstach naukowych występuje uniwersalny angielski, praktycznie nieobecne są neologizmy i wyrażenia slangowe. Wysiłek, jaki trzeba włożyć, żeby sprawnie przyswajać naukowe teksty anglojęzyczne, jest nieporównywalnie mniejszy, niż gdyby się chciało równie łatwo czytać poezję lub prozę w tym języku. Będzie on procentował przy wielu okazjach, bo jest to podstawowy język globalnej komunikacji, nie tylko naukowej.

Czy korzystanie z dobrych polskich artykułów i monografii, ewentualnie regularny udział w kursach i szkoleniach dostępnych w języku polskim nie jest wystarczający? W moim przekonaniu nie jest. Źródła te nierzadko przedstawiają wyniki badań w sposób uproszczony, zerojedynkowy. Największym jednak ich ograniczeniem jest to, że przedstawiają tylko ułamek wartościowej wiedzy. Jako pracownik naukowy wiem, że w obiegu naukowym funkcjonuje wiele wartościowych idei, które nie są omówione w publikacjach w języku polskim (jest to niejednokrotnie problem przy układaniu programu zajęć na studiach psychologicznych). Tylko niewielka część wartościowej wiedzy zostanie w tej czy innej formie spolszczona, często po bardzo wielu latach. Oczywiście, teksty polskie niejednokrotnie stanowią bardzo cenne źródło wiedzy, ale ograniczając się tylko do nich, niejako siłą rzeczy akceptujemy to, że nasza wiedza będzie wyrywkowa i będzie zależała od tego, czy autor tekstu zapoznał się z najnowszymi osiągnięciami w danej dziedzinie nauki.

Nie twierdzę też, że każde twierdzenie, które znajdziemy w artykule przeglądowym (albo nawet popularnonaukowym), musi być przez nas zweryfikowane w podstawowych źródłach. Nikt nie ma tyle czasu, żeby mógł każdą jednostkę informacji szczegółowo sprawdzić. Szczególnie trudno tego wymagać, jeśli konkretny fragment wiedzy nie jest szczególnie istotny w naszej pracy zawodowej. Ale, przykładowo, jeśli chcemy pracować z ludźmi cierpiącymi na zaburzenia psychiczne przy pomocy techniki, o której przeczytaliśmy wzmiankę w popularnym czasopiśmie, jest głęboko nieprofesjonalne, żeby nie sięgnąć do podstawowych źródeł. Po ich lekturze często okazuje się, że skuteczność tej techniki jest, obiektywnie patrząc, dosyć słaba, albo też samo badanie jest wykonane na jakiejś szczególnej grupie, lub ma wątpliwą metodologię.

Mam wielką nadzieję, że tym wstępem przekonałem Czytelnika do tego, że czytanie podstawowych źródeł (czyli artykułów empirycznych w czasopismach naukowych) jest elementem pracy profesjonalnego psychologa. Przejdę zatem do właściwej części artykułu, która jest podzielona na trzy sekcje, mianowicie dotyczące interpretowania istotności statystycznej, wielkości efektu i przedziałów ufności.

INTERPRETACJA ISTOTNOŚCI STATYSTYCZNEJ (I KRYTYCZNE UWAGI WOBEC TEJ MIARY)

Praktycznie wszystkie współczesne artykuły empiryczne mają jeden wspólny mianownik: wskaźnik istotności statystycznej badanych efektów. Według analizy R. Hubbarda i P.A. Ryana (2000) mniej więcej od lat 50. XX wieku ponad 90% artykułów zawiera ten wskaźnik. Istnieje wiele testów statystycznych, różnych w zależności od rodzaju danych i planu badawczego. Ich interpretacja bywa trudna, bo są wyliczane według innych zasad, a dodatkowo ich wielkość jest uzależniona od liczby zbadanych osób. Tymczasem wskaźnik istotności statystycznej ma tę zaletę, że niezależnie od rodzaju testu i wielkości próby jest interpretowany w ten sam sposób. Informuje nas, jakie jest prawdopodobieństwo, że otrzymalibyśmy określony wynik, gdyby hipoteza zerowa była prawdziwa. Co to jest hipoteza zerowa? Mówiąc ogólnie, to w nauce odpowiednik domniemania niewinności w prawie. Polega ona na przyjęciu założenia, że nie ma żadnych zależności (korelacji, różnic między średnimi itd.), dopóki nie ma mocnych dowodów, że jest inaczej.

Przeanalizujmy kilka przykładów. Wyobraźmy sobie, że mierzymy wzrost reprezentatywnej próby tysiąca mieszkańców Szczecina i Przemyśla. Nie mamy żadnych podstaw, żeby podejrzewać, że grupy te różnią się znacząco wzrostem, ale jesteśmy otwarci na to, co przyniesie badanie. Formułujemy hipotezę zerową, czyli przyjmujemy, że różnica średnich wzrostów między mieszkańcami dwóch miast wynosi 0 cm. Jest jednak skrajnie mało prawdopodobne, że grupy te uzyskują wynik identyczny do dziesiątego miejsca po przecinku. A więc praktycznie zawsze hipoteza zerowa traktowana literalnie jest fałszywa. Jednak nieduże różnice mogą być wynikiem przypadku i dzięki wskaźnikowi istotności statystycznej powinniśmy to uchwycić.

Załóżmy dalej, że wynik naszego badania jest taki, że szczecinianie są wyżsi o 0,3 cm. Dzięki odpowiednim testom statystycznym możemy wyliczyć wskaźnik istotności statystycznej. Jak już wspomniałem, określa on prawdopodobieństwo takiego wyniku, gdyby hipoteza zerowa była prawdziwa, a mówiąc precyzyjniej, jakie jest prawdopodobieństwo tego wyniku i wszystkich wyników bardziej skrajnych (bo jeżeli hipoteza zerowa jest fałszywa dla 0,3 cm, to tym bardziej jest fałszywa dla różnicy rzędu 5 cm). Zwykle przy wyliczaniu istotności statystycznej zakłada się istotność dwustronną, tj. w naszym przykładzie wylicza się dodatkowo, jakie jest prawdopodobieństwo, że to przemyślanie są wyżsi o 0,3 cm lub więcej.

Załóżmy dalej, że przy pomocy odpowiednich testów statystycznych obliczono, że istotność statystyczna różnicy 0,3 cm wynosi 11,4%. Przyjęła się konwencja, żeby poziom istotności oznaczać literą *p* i zapisywać jako ułamek dziesiętny. Zatem w naszym przykładzie p = 0,114 (ważna uwaga techniczna: w artykułach często jedną, dwoma

i trzema gwiazdkami [*, **, ***] oznacza się poziomy istotności statystycznej odpowiednio mniejsze od 0,05, 0,01 i 0,001).

Czy w naszym przykładzie mieszkańcy Szczecina i Przemyśla różnią się wzrostem? Zgodnie z zasadami testowania istotności różnic nie różnią się, ponieważ wynik jest wtedy istotny statystycznie, jeśli wskaźnik istotności jest mniejszy niż 5% (a więc $p < 0,05$). W omawianym tu przypadku taki nie jest. Dlaczego dane osiągają poziom istotności przy mniej niż 5%, a nie na przykład poniżej 4,65%? Nie wynika to z żadnego wzoru ani prawa statystycznego. Po prostu taką przyjęto konwencję.

Istnieje wiele problemów związanych z istotnością statystyczną. Bywa ona interpretowana jako istotność „w ogóle", tj. informacja, że wynik jest znaczący. Niekoniecznie tak musi być, gdyż wskaźnik istotności statystycznej uzależniony jest też od wielkości próby. Przykładowo, gdybyśmy zmierzyli wzrost miliona mieszkańców Warszawy i miliona mieszkańców Londynu, to mikroskopijne różnice (około 0,2 mm) okazałyby się istotne statystycznie. Jednak czujemy intuicyjnie, że różnice takie nie mają żadnego znaczenia praktycznego.

Żeby nie być gołosłownym podam pewien przykład, dobrze obrazujący omawiany problem. W numerze czasopisma popularnonaukowego „Charaktery" dowiedziałem się o ciekawych badaniach dotyczących związku osobowości z preferencjami artystycznymi (Gelitz, 2011). W tym artykule można przeczytać, że „osoby lubiące obrazy impresjonistów są sumienne i raczej zamknięte na nowe doświadczenia" (Gelitz, 2011, s. 35). Z ciekawości sprawdziłem te dane w oryginalnym źródle (czyli w artykule T. Chamorro-Premuzic i in., 2009). Okazało się, że owszem osoby sumienne preferują obrazy impresjonistów, ale siła tego efektu jest bardzo słaba ($r = 0,08$; interpretacja takich wielkości w następnej sekcji). Związek otwartości na doświadczenie z preferowaniem impresjonizmu jest wręcz mikroskopijny ($r = 0,01$). Przebadano jednak w badaniu internetowym bardzo dużą próbę, ponad 90 tysięcy osób, stąd wszystkie korelacje różne od 0,00 były istotne statystycznie.

Istotność statystyczna bywa też błędnie interpretowana, jako informacja, z jakim prawdopodobieństwem uda się określone badanie powtórzyć. Stąd jeśli $p = 0,05$, to niektórzy błędnie przyjmują, że jest 95% szans, że uda się powtórzyć wynik badania, jest to jednak błąd (por. Gigerenzer, 2004). Dodatkowo rozrzut otrzymywanych empirycznie wartości istotności statystycznej jest bardzo duży i w niewielkim stopniu pozwala nam ona wnioskować o tym, czy w następnym badaniu otrzymamy istotnie statystyczny efekt (chyba, że wartość p jest naprawdę bardzo niska, przykładowo $p < 0,001$; por. Cumming, 2013, rodz. 5).

Do sceptycyzmu i ostrożności zmuszają też pojawiające się w ostatnich latach doniesienia, że badacze bardzo często stosują różne sztuczki statystyczne, żeby osiągnąć pożądane $p < 0,05$ (John, Loewenstein, Prelec, 2012). Być może dlatego w literaturze

jest obecnie nieprawdopodobnie dużo wartości p nieznacznie tylko mniejszych od 0,05 (por. Masicampo, Lalande, 2012).

W rzeczy samej we współczesnej psychologii pojawia się coraz więcej głosów, żeby wycofać się z podawania istotności statystycznej albo używać jej tylko jako dodatkowy wskaźnik (Cohen, 1994/2006; Cumming, 2013; Kline, 2013). Jest ona jednak tak rozpowszechniona, że prawdopodobnie zajmie długie lata jej wyrugowanie z praktyk publikacyjnych (o ile w ogóle). Nie zniknie też z setek tysięcy już opublikowanych artykułów. Stąd Czytelnik powinien wiedzieć, jak ją zinterpretować, mając jednak świadomość jej niedoskonałości.

INTERPRETACJA WIELKOŚCI EFEKTÓW

Na początek załóżmy, że testujemy empirycznie nową terapię wspomagającą rzucenie palenia. Okazało się, że po trzymiesięcznej terapii z paleniem rozstało się 50% uczestników terapii. Wiemy więc, że skuteczność określonej terapii jest niemała, choć będzie nadużyciem reklamować ją jako cudowny i niezawodny dla każdego sposób na rozstanie się z nałogiem.

Jaka jest istotność statystyczna tego efektu? Sam odsetek osób, które skutecznie rzuciły palenie, nam tego nie powie. Wskaźnik istotności statystycznej będzie różny w zależności od tego, jak dużą grupę zbadaliśmy. Tym niemniej i bez niego informacja o tym, że 50% osób porzuciło palenie, jest intuicyjnie uchwytna i zrozumiała. Owo „50%" jest najprostszym przykładem wielkości efektu. Wielkości efektu to ilościowa miara natężenia określonego zjawiska będącego przedmiotem zainteresowania badaczy.

Wyobraźmy sobie teraz, że dowiadujemy się, że po zastosowaniu pewnej nowatorskiej terapii w grupie palaczy średnia liczba wypalanych dziennie papierosów zmniejszyła się o osiem. Różnica średnich też jest rodzajem miary wielkości efektu. Czy osiem papierosów to mało, czy dużo? Choć jednostka jest intuicyjnie łatwo zrozumiała, jest to już bardziej problematyczne w interpretacji. Przykładowo, nie wiemy, czy badani palili wcześniej mało, czy dużo. Te osiem papierosów mniej dziennie może oznaczać, że prawie wszyscy rzucili palenie albo że badani nieznacznie tylko ograniczyli nałóg. Mówiąc inaczej, podana różnica średnich to mało lub dużo w zależności od tego, jaka jest ogólna zmienność (wariancja) danych.

W celu lepszego zobrazowania zagadnienia podany zostanie jeszcze inny przykład. Załóżmy, że okazuje się, że dzięki nowatorskiej technice motywowania biegacze poprawili swój czas o 1 sekundę. Wiele osób pewnie nie zadowoli się tą informacją, ale intuicyjnie zapyta „a na jakim dystansie?". Jeśli mówimy o 1 sekundzie w maratonie, poprawa taka praktycznie nic nie znaczy. Natomiast 1 sekunda na 100 metrów to gi-

gantyczny postęp. A więc określony wynik może znaczyć mało lub dużo w zależności od ogólnej zmienności danych.

Dodatkowo w psychologii wiele danych jest wyrażonych w trudno interpretowalnych jednostkach. Przykładowo, czy to dużo czy mało, jeśli księgowi w porównaniu z artystami uzyskują 34 punkty więcej w teście mierzącym potrzebę domknięcia poznawczego (Webster, Kruglanski, 1994)? Dlatego wypracowano w statystyce miary wielkości efektów, które uwzględniają ogólną zmienność danych i które mogą być stosowane do dowolnych jednostek. O takich miarach mówimy, że są wystandaryzowane, co oznacza, że wszelkie dane można wyrazić przy pomocy tych miar i określone wielkości interpretować w ten sam sposób.

Najpopularniejszą z takich miar jest d Cohena. Obliczana jest ona w ten sposób, że różnica średnich (dwóch grup, np. eksperymentalnej i kontrolnej) dzielona jest przez miarę zmienności danych, konkretnie wielkość odchylenia standardowego. Jej wielkość rozciąga się od zera do teoretycznie nieskończoności (możliwe są też wyniki ujemne, ale one oznaczają po prostu inny kierunek zależności), jednak w psychologii rzadko spotykamy wielkości większe niż 1,0 i ekstremalnie rzadko większe niż 2,0.

Przykładowo, metaanalizy podają, że skuteczność psychoterapii poznawczo-behawioralnej w leczeniu depresji wynosi około d = 0,8 w porównaniu z grupą osób niepoddanych żadnej terapii (np. Gloaguen i in., 1998). Oznacza to literalnie, że osoby po psychoterapii uzyskują na określonej skali zmiennych zależnych (np. w skali depresji Becka) wyniki o 0,8 odchylenia standardowego lepsze.

Określona liczba odchyleń standardowych jest trudna do interpretacji dla niestatystyków. Czy można łatwiej zinterpretować konkretną wielkość d Cohena? W tab. 1 podano, jakie wartości umownie uznaje się za małe, średnie, duże i bardzo duże (por. Cohen, 1988 i Rosenthal, 1996). Istnieją możliwości jeszcze bardziej precyzyjnej interpretacji, które zostały przedstawione w tab. 2. W pierwszej kolumnie podano konkretne wielkości d Cohena, a w drugiej oznaczono, jaki procent osób (z grupy o niższej średniej) ma wynik poniżej przeciętnego wyniku osoby z grupy o wyższej średniej. Jeśli przykładowo różnica w dobrostanie po terapii wynosi d = 0,8, oznacza to, że średnio szczęśliwa osoba po terapii jest bardziej szczęśliwa niż 79% osób bez terapii. Gdyby wynik wynosił -0,8, oznaczałoby to, że średnio szczęśliwa osoba po terapii ma dobrostan gorszy niż 79% osób bez terapii (a lepszy niż 21%). Ostatnia kolumna oznacza prawdopodobieństwo, z jakim losowo wybrana osoba z grupy (o wyższym wyniku) będzie miała wyższy wynik niż osoba z drugiej grupy. A więc zgodnie z naszym przykładem: jeśli terapia w stosunku do braku terapii jest skuteczna na poziomie d = 0,8, to jeśli wylosujemy z populacji osób z zaburzeniami jedną osobę po terapii i jednego pechowca bez terapii, to istnieje 71% szansy, że osoba po terapii będzie miała wyższy wynik.

Tab. 1. Podstawowe miary wielkości efektu, miary, które są interpretowane podobnie, oraz orientacyjne wielkości ułatwiające interpretacje[1]

Podstawowe miary wielkości efektu	Miary siły efektu, których wielkość interpretujemy w podobny sposób	Efekty słabe	Efekty średnie	Efekty silne	Efekty bardzo silne
d (Cohena)	Δ, d Glassa, g Hedgesa	0,2-0,5	0,5-0,8	0,8-1,3	> 1,3
r (Pearsona)	R, φ, ρ, β, τ (tau) oraz ich pochodne*	0,1-0,3	0,3-0,5	0,5-0,7	> 0,7
r*	R^2, η^2, ω^2, ε^2	0,01-0,09	0,09-0,25	0,25-0,49	> 0,49

Źródło: na podstawie J. Cohen (1988) i J.A. Rosenthal (1996).

*Pochodne tych wielkości efektów są oznaczane dodatkowymi symbolami, np. R_{adj}, ω_p itp.

Statystykę d Cohena stosuje się, gdy porównujemy wyniki osób z dwóch grup (np. kobiet i mężczyzn, chorych i zdrowych, osób przed terapią i po terapii, grupy kontrolnej i eksperymentalnej). W tym wypadku zmienna niezależna („przyczyna") jest kategorialna (inaczej nominalna), a zmienna zależna („skutek") jest ilościowa. Nierzadko jednak w psychologii mamy do czynienia z sytuacją, gdy chcemy porównać ze sobą dwie zmienne ilościowe. Przykładowo chcemy sprawdzić, jaka jest siła związku między inteligencją a zarobkami. W takim wypadku najczęściej używaną statystyką jest r (Pearsona) i jest to miara korelacji. Interpretujemy ją inaczej niż wielkość d Cohena. Po pierwsze wartość bezwzględna tej miary musi być w zakresie <0,1>, gdzie 0 to brak związku, a 1 to najsilniejszy możliwy związek (tak jak w przypadku d Cohena znowu możliwe są ujemne korelacje i analogicznie nie oznaczają one związku słabszego niż zero, ale po prostu inny kierunek zależności).

Jak interpretować konkretne wartości r Pearsona? W literaturze przyjmuje się umownie, że wielkości podane w tab. 1 są odpowiednio: małe, średnie, duże i bardzo duże. W analizie G.E. Gignaca i E.T. Szodorai (2016) dotyczącej spotykanych wielkości efektów w literaturze różnic indywidualnych stwierdzono jednak, że wielkość korelacji $r > 0,5$ występuje tylko w 3% opisywanych w literaturze naukowej efektów. Z kolei 75% korelacji ma wartość bezwzględną $r < 0,29$, a połowa jest słabsza od 0,19. Zatem być może umowne wielkości efektów są zbyt wyśrubowane relatywnie w stosunku do tego, co odkrywamy w badaniach psychologicznych.

Miarę korelacji można podnieść do kwadratu i wtedy wielkość r^2 informuje nas o tym, ile procent wariancji (czyli zmienności) udało się wyjaśnić. Przykładowo korelacja między sumiennością a osiągnięciami akademickimi to około $r = 0,24$ (O'Connor, Paunonen, 2007), a więc po podniesieniu do kwadratu $r^2 = 0,057$. Oznacza to, że

[1] Wymowa greckich liter jest następująca: Δ – delta; φ – fi, ρ – rho; β – beta, τ – tau; η^2 – eta-kwadrat, ω^2 – omega kwadrat, ε^2 – epsilon-kwadrat (uwaga: nie należy mylić współczynnika korelacji ρ (rho) ze współczynnikiem istotności statystycznej p).

Tab. 2. Sposoby interpretowania wielkości efektu d Cohena

Wielkość efektu (d Cohena)	Odsetek osób z grupy kontrolnej*, które mają wynik niższy niż średni wynik w grupie eksperymentalnej	Prawdopodobieństwo, że osoba z grupy eksperymentalnej będzie miała wyższy wynik niż osoba z grupy kontrolnej, jeśli wybierzemy z populacji dwie osoby w sposób losowy
0.0	50%	0,50
0.1	54%	0,53
0.2	58%	0,56
0.3	62%	0,58
0.4	66%	0,61
0.5	69%	0,64
0.6	73%	0,66
0.7	76%	0,69
0.8	79%	0,71
0.9	82%	0,74
1.0	84%	0,76
1.2	88%	0,80
1.4	92%	0,84
1.6	95%	0,87
1.8	96%	0,90
2.0	98%	0,92

Źródło: wyliczono na podstawie wzorów z pracy K.O. McGraw i S.P. Wong (1992).

*Zakładamy, że osoba z grupy eksperymentalnej ma wyższy wynik. Wartości ujemne d Cohena interpretowalibyśmy analogicznie, ale w przeciwnym kierunku. Oczywiście d Cohena możemy stosować nie tylko do porównań grup eksperymentalnych, lecz także grup, takich jak płeć, choroba/brak choroby, osoby przed terapią i po terapii itd.

sumienność wyjaśnia[2] 5,7% zmienności sukcesu akademickiego. Nie jest to bardzo dużo, ale z drugiej strony – nie licząc inteligencji ogólnej – na tle innych zmiennych dyspozycyjnych jest to jeden z najlepszych prognostyków.

Jest jeszcze inny sposób interpretacji. W tab. 3 obok wielkości współczynnika korelacji r podano w drugiej kolumnie liczbę, którą interpretujemy następująco. Załóżmy, że mamy korelacje dwóch zmiennych (np. inteligencja i zarobki). Wyznaczmy mediany rozkładów tych dwóch zmiennych. Jeśli korelacja pomiędzy zmiennymi wynosi r = 0,00, to wśród osób, które mają inteligencję powyżej mediany, 50% osób będzie

[2] Przy założeniu, że jest to zależność przyczynowo-skutkowa, a nie pozorna. Badania korelacyjne nie udzielają na to jednoznacznej odpowiedzi.

Tab. 3. Interpretacja szczegółowa wielkości korelacji r Pearsona

Wielkość współczynnika korelacji r (Pearsona)	Odsetek, który mając jedną zmienną powyżej mediany, ma równocześnie wynik w drugiej zmiennej powyżej mediany
0,00	50,0%
0,05	51,5%
0,10	53,1%
0,15	54,7%
0,20	56,4%
0,25	58,0%
0,30	59,7%
0,35	61,3%
0,40	63,1%
0,45	64,8%
0,50	66,6%
0,55	68,5%
0,60	70,4%
0,65	72,5%
0,70	74,6%
0,75	76,9%
0,80	79,5%
0,85	82,3%
0,90	85,6%
0,95	89,8%
1,00	100,0%

Źródło: opracowanie na podstawie W.B. Michael (1966).

miało zarobki powyżej mediany i 50% osób będzie miało zarobki poniżej mediany. Jeśli jednak korelacja ta jest niezerowa i wynosi przykładowo r = 0,30 (tyle rzeczywiście mniej więcej wynosi, np. Zagorsky, 2007), to wśród osób, które mają inteligencję powyżej mediany, 59,7% osób będzie miało zarobki powyżej mediany, a 40,3% osób będzie miało zarobki poniżej mediany. Na pierwszy rzut oka różnice nie są bardzo duże, ale wśród osób o podwyższonej inteligencji jest prawie 50% więcej osób o podwyższonych zarobkach, jeśli za punkt odniesienia przyjąć grupę osób o inteligencji poniżej mediany. Gdyby to była korelacja ujemna, przykładowo także r = -0,30, interpretacja byłaby odwrotna (jeśli ktoś ma inteligencję powyżej mediany, to ma 40,3% szans, że ma zarobki poniżej mediany).

Miary wielkości efektu d Cohena, r Pearsona (oraz jej pochodna r^2) są najczęściej występującymi miarami wielkości efektu spotykanymi w literaturze. Interpretacja ich wielkości jest inna, ale możemy łatwo przekształcić jedną w drugą przy pomocy wzorów.

$$d = 2r/[(1 - r^2)^{1/2}]$$

$$r = [d^2/(d^2 + 4)]^{1/2}$$

Istnieje też wiele innych statystyk, które są konceptualnie do nich zbliżone i które stosujemy do innych rodzajów danych, ewentualnie przy ich obliczaniu przyjmujemy trochę inne założenia. Ich szczegółowe omówienie matematyczno-metodologiczne wymagałoby znacznie szerszego artykułu. W tym momencie wystarczy nam wiedza, że interpretacja ich wielkości jest zbliżona. W tab. 1 w drugiej kolumnie wyliczyłem te miary wielkości efektów. Dodatkowo w tab. 4 przedstawiłem te statystyki, których wielkości nie należy interpretować. Pojawiają się one regularnie w artykułach, ale tylko dlatego, że są potrzebne do wyliczenia poziomów istotności statystycznej. Najczęściej spotykaną statystyką jest t (Studenta). Jeśli znamy jego wartość oraz liczbę stopni swobody (oznaczane jako df; ang. *degrees of freedom*), to możemy wyliczyć d Cohena za pomocą wzoru[3].

$$d = 2t / (df^{1/2})$$

Tab. 4. Zestawienie statystyk, których wielkości nie powinno się interpretować, gdyż są zależne od wielkości próby

Statystyka	Kiedy jest stosowana
t (Studenta)	Porównywanie średnich przy danych ilościowych
U (Manna-Whitneya), W (Wilcoxona)	Porównywanie średnich (rang) przy danych porządkowych
$\chi 2$ (chi-kwadrat)	Tabele krzyżowe (zmienne nominalne x zmienne nominalne)
F	Analiza wariancji

Źródło: opracowanie własne.

Poza wyżej wymienionymi Czytelnik może jeszcze spotkać miary efektu oparte na ryzyku, najczęściej spotykane to ryzyko względne (RR, ang. *relative risk*) i iloraz szans (OR, ang. *odds ratio*). Dane takie występują dużo częściej w czasopismach medycznych, niemniej sporadycznie pojawiają się w psychologii (np. Walker i in., 2002 badali ryzyko względne leczenia psychiatrycznego w zależności od ilorazu inteligencji).

[3] Alternatywnie możemy zamiast df wstawić N – 2, gdzie N to wielkość próby badanej.

Przykładowo w jednym z pierwszych badań sprawdzających skuteczność szczepionki na polio wyliczono, że w grupie prawie 200 tysięcy zaszczepionych osób polio rozwinęło się u 33, natomiast w podobnie licznej grupie niezaszczepionych zachorowało 115 (dane za: Rosnow, Rosenthal, 2003). Gdyby na podstawie tych danych obliczyć standardowe wielkości efektu, to okazałoby się, że są one mikroskopijne (większość osób tak czy owak nie zachorowała na polio). Jednak w tym wypadku jest to mało sensowne, gdyż możemy założyć, że olbrzymia większość osób w badaniu nie zetknęła się z wirusem polio i nie miała szans na rozwinięcie choroby. Jeśli natomiast porównamy ryzyko względne, to okaże się, że szansa zapadnięcia na polio w grupie zaszczepionej jest prawie 3,5 razy mniejsza niż w grupie kontrolnej.

Przy interpretowaniu zarówno ryzyka względnego, jak i ilorazu szans wskazuje się, że wielkość 2,0 to słaby efekt, 3,0 to efekt umiarkowany, a 4,0 to efekt silny (Ferguson, 2009). Im większa bazowa zachorowalność, tym określone ryzyko względne nabiera większego znaczenia.

Na koniec tej sekcji chciałbym podać dwie przestrogi. Przedstawione tu wielkości efektów określane jako duże, średnie i małe należy traktować z ostrożnością. Ostateczna interpretacja danej wielkości efektu zależy od tego, jaki efekt osiągamy oraz od tego, jakie koszty ponosimy, żeby osiągnąć ten efekt. Przykładowo, efekty polegające na obniżeniu śmiertelności lub znaczącej poprawie zdrowia są zawsze warte uwagi, nawet jeśli są bardzo słabe. Jeśli, na przykład, dowiemy się, że pewna interwencja psychologiczna powoduje, że 1% osób jej poddanych trwale przestaje palić, to statystycznie patrząc, jest to bardzo słaby efekt. Biorąc jednak pod uwagę, że palenie ma jednoznacznie szkodliwe skutki, to nawet obniżenie odsetka palących o 1% jest wysoce pożądane. Jeśli dodatkowo okazuje się, że do uzyskania takiego efektu potrzebna jest dwuminutowa rozmowa, to mimo wszystko efekt ten możemy uznać za znaczący, ponieważ niewielki wysiłek prowadzi do bardzo pozytywnych skutków. Gdyby się okazało w innym badaniu, że wielomiesięczne kampanie reklamowe kosztujące kilkadziesiąt milionów złotych również obniżają odsetek palących o 1% (wśród tych, którzy regularnie te reklamy widywali), to choć efekt ten jest – statystycznie patrząc – równie silny, to jednak w szerszym kontekście nie wygląda już tak imponująco.

Warto również pamiętać o tym, że małe efekty mogą się kumulować. Jeśli jakieś szkolenie psychologiczne podnosi efektywność sprzedaży o 2%, to na krótką metę nie będzie to spektakularnie dużo, ale w ciągu lat może się przełożyć na znaczące zyski.

INTERPRETACJA PRZEDZIAŁÓW UFNOŚCI

Jak już wspominałem, w kontrze do stosowania testów istotności statystycznej i zerojedynkowego odrzucania lub przyjmowania hipotezy badawczej, coraz większego

znaczenia nabiera sposób myślenia, który zakłada, że podstawowymi statystykami omawianymi w artykułach empirycznych powinny być wielkości efektu i przedziały ufności.

Pojawia się jednak pytanie o to, dlaczego same wielkości efektu nie są wystarczające? W badaniach empirycznych występuje zjawisko nazywane błędem próbkowania. Z zupełnie przypadkowych względów wynik naszej próby może odbiegać od rzeczywistych wyników populacji, czyli tak zwanego wyniku prawdziwego. Podkreślę, że mówimy tu tylko o wpływie przypadku, a nie o systematycznych zniekształceniach wyników wynikających z błędów metodologicznych lub świadomego oszustwa.

Przedziały ufności są to statystyki, które pozwalają określić, gdzie z dużym prawdopodobieństwem znajduje się wynik prawdziwy. Przedziały ufności można wyznaczyć dla wszelkich typów wielkości efektów (odsetka z populacji, różnicy średnich, wielkości korelacji itd.). Standardowo przedziały ufności oznacza się literkami CI (ang. *confidence interwal*), podaje określoną wartość przedziału ufności oraz górny i dolny wynik dla przyjętej wartości. Czasami możemy też znaleźć oznaczenie dolnej i górnej wartości przedziału ufności, opisane odpowiednio jako LL i UL (od ang: *lower limit* i *upper limit*).

Przykładowo – w artykule możemy znaleźć takie wyrażenie: „wielkość korelacji wynosi r = 0,04 (CI 95% = [-0,131; 0,209])". Jak interpretować tę konkretną informację? W 95 przypadkach na 100 w przedziale <-0,131; 0,209> mieści się rzeczywista (tj. występująca w populacji) wartość korelacji. W naszym przykładzie może ona być zarówno ujemna, jak i dodatnia, choć raczej jest słaba. Najbardziej prawdopodobne jest, że rzeczywista wielkość korelacji jest zbliżona do 0,04, a przynajmniej taka wartość jest bardziej prawdopodobna niż skrajny wynik (np. 0,209). Jest techniczną nieścisłością zinterpretować przedział ufności jako informację, że wynik prawdziwy jest na 95% w tym przedziale. Wyznaczony przez nas przedział ufności może go obejmować, ale nie musi, więc szansa na to wynosi albo 100%, albo 0%. Jednak w serii wielu badań w 95% przypadków przedziały ufności wskażą wynik prawdziwy.

Przedziały ufności dają nam jeszcze dodatkową pożyteczną informację. Pozwalają mianowicie określić precyzję naszego badania. Jeśli otrzymujemy informację, że „wielkość korelacji wynosi r = 0,51 (CI 95% = [-0,09; 0,83])", dowiadujemy się, że otrzymaliśmy silny efekt, ale precyzja naszego badania jest bardzo niska. Jest możliwe, że nasz efekt jest zerowy czy wręcz ujemny (choć oczywiście może też być jeszcze silniejszy).

KILKA UWAG NA PODSUMOWANIE

Podkreślę na koniec, że interpretacja badań empirycznych musi być zawsze dużo szersza niż tylko analiza samych statystyk. W badaniach mogą występować rozmaite

artefakty metodologiczne, przez które możemy uzyskiwać silne efekty niewystępujące w rzeczywistości (te artefakty to m.in. efekt oczekiwań interpersonalnych, zgadywanie hipotezy przez badanego, uwrażliwianie przez pomiar, niereprezentatywność prób itd.; wyczerpujące omówienie na ten temat zawiera podręcznik J. Brzezińskiego [2012]). Sporadycznie, wyniki niektórych badań są po prostu oszustwem, co dobitnie pokazał kilka lat temu przykład wybitnego wówczas psychologa społecznego D. Stapela. Dlatego, aby uzyskać możliwie najbardziej wyczerpującą wiedzę, konieczne jest analizowanie więcej niż wyników jednego badania. Jedno z podstawowych pytań, jakie powinniśmy sobie stawiać, to „czy ktoś powtórzył to badanie". Zadając konkretne pytanie badawcze, powinniśmy szukać wielu badań lub w miarę możliwości metaanaliz. Metaanalizy są szczególnym typem artykułu empirycznego, który zawiera ilościowe podsumowanie wielu badań. Metaanalizy też czasami podają mylne informacje, ale są najpewniejszym źródłem wiedzy.

LITERATURA

Brzeziński, J. (2012). *Metodologia badań psychologicznych*. Warszawa: Wydawnictwo Naukowe PWN.

Chamorro-Premuzic, T., Reimers, S., Hsu, A., Ahmetoglu, G. (2009). Who art thou? Personality predictors of artistic preferences in a large UK sample: The importance of openness. *British Journal of Psychology*, *100*(3), 501-516.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale: Lawrence Erlbaum.

Cohen, J. (1994/2006). Ziemia jest okrągła (p < 0,05). W: J. Brzeziński, J. Siuta (red.), *Metodologiczne i statystyczne problemy psychologii* (100-118). Poznań: Zysk i S-ka.

Cumming, G. (2013). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York: Routledge.

Ferguson, C.J. (2009). An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research and Practice*, *40*(5), 532-538.

Francuz, P., Mackiewicz, R. (2007). *Liczby nie wiedzą, skąd pochodzą. Przewodnik po metodologii i statystyce nie tylko dla psychologów*. Lublin: Wydawnictwo KUL.

Gelitz, Ch. (2011). Podyskutujmy o gustach. *Charaktery*, *3*, 34-38.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, *33*, 587-606.

Gignac, G.E., Szodorai, E.T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences*, *102*, 74-78.

Gloaguen, V., Cottraux, J., Cucherat, M., Blackburn, I. (1998). A meta-analysis of the effects of cognitive therapy in depressed patients. *Journal of Affective Disorders*, *49*, 59-72.

Hubbard, R., Ryan, P.A. (2000). The historical growth of statistical significance testing in psychology – And its future prospects. *Educational and Psychological Measurement*, *60*(5), 661-681.

John, L., Loewenstein, G., Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth-telling. *Psychological Science*, *23*, 524-532.

King, B., Minium, E. (2009). *Statystyka dla psychologów i pedagogów.* Warszawa: Wydawnictwo Naukowe PWN.

Kline, R.B. (2013). *Beyond significance testing. Statistics reform in the behavioral sciences*. American Psychological Association.
Masicampo, E.J., Lalande, D.R. (2012). A peculiar prevalence of p values just below .05. *The Quarterly Journal of Experimental Psychology*, *65*(11), 2271-2279.
Michael, W.B. (1966). An Interpretation of the Coefficients of Predictive Validity and of Determination in Terms of the Proportions of Correct Inclusions or Exclusions in Cells of a Fourfold Table. *Educational and Psychological Measurement*, *26*(2), 419-426.
McGraw, K.O., Wong, S.P. (1992). A common language effect size statistic. *Psychological Bulletin*, *111*(2), 361-365.
O'Connor, M.C., Paunonen, S.V. (2007). Big Five personality predictors of post-secondary academic performance. *Personality and Individual Differences*, *43*(5), 971-990.
Rosenthal, J.A. (1996). Qualitative descriptors of strength of association and effect size. *Journal of Social Service Research*, *21*(4), 37-59.
Rosnow, R.L., Rosenthal, R. (2003). Effect sizes for experimenting psychologists. *Canadian Journal of Experimental Psychology*, *57*(3), 221.
Walker, N.P., McConville, P.M., Hunter, D., Deary, I.J., Whalley, L.J. (2002). Childhood mental ability and lifetime psychiatric contact: A 66-year follow-up study of the 1932 Scottish Mental Ability Survey. *Intelligence*, *30*(3), 233-245.
Webster, D.M., Kruglanski, A.W. (1994). Individual differences in need for cognitive closure. *Journal of Personality and Social Psychology*, *67*(6), 1049-1062.
Zagorsky, J.L. (2007). Do you have to be smart to be rich? The impact of IQ on wealth, income and financial distress. *Intelligence*, *35*(5), 489-501.

INTERPRETACJA STATYSTYK W ARTYKUŁACH NAUKOWYCH – WSKAZÓWKI DLA PRAKTYKÓW

STRESZCZENIE: Artykuł zawiera informacje o tym, jak interpretować podstawowe dane statystyczne: wskaźniki istotności statystycznej, wielkości efektu i przedziały ufności. Pokazano kilka heurystyk użytecznych przy interpretacji wielkości efektów korelacji r Pearsona, statystyki d Cohena oraz relatywnego ryzyka. Olbrzymia większość pozostałych efektów jest pochodną wyżej wymienionych. Dodatkowo wskazano również, jakie są ograniczenia wybranych wskaźników, szczególnie istotności statystycznej. Artykuł jest pomyślany jako pomoc szczególnie dla psychologów praktyków.
SŁOWA KLUCZOWE: wielkość efektu, istotność statystyczna, przedziały ufności, interpretowanie danych statystycznych.

THE INTERPRETATION OF THE STATISTICAL DATA IN SCIENTIFIC PAPERS – ADVICES FOR PRACTITIONERS

SUMMARY: The article contains information how to interpret statistical data: statistical significance, effect size and confidence intervals. Several heuristics are given how to usefully interpret the magnitude of the correlation Pearson's r, Cohen's d and relative risk. The vast majority of other effects is a derivative of the aforementioned. In addition, I also show the limitations of selected indicators, especially statistical significance. This article is intended as an aid especially for psychologists practitioners.
KEYWORDS: effect size, statistical significance, confidence interval, interpretation of statistical data.