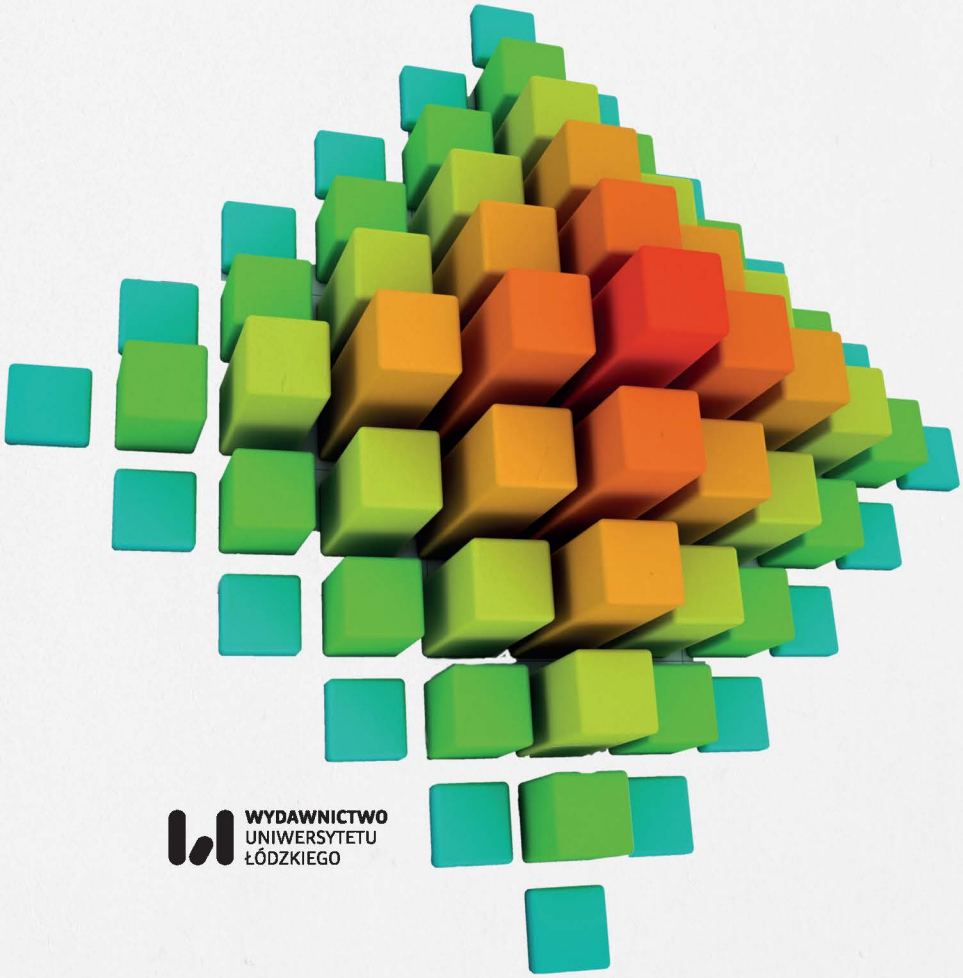
Wiesław Szymczak

Praktyka wnioskowania statystycznego



WYDAWNICTWO
UNIWERSYTETU
ŁÓDZKIEGO

Praktyka wnioskowania statystycznego

WYDAWNICTWO
UNIWERSYTETU
ŁÓDZKIEGO

Wiesław Szymczak

Praktyka wnioskowania statystycznego

WYDAWNICTWO
UNIWERSYTETU
ŁÓDZKIEGO

Łódź 2018

Wiesław Szymczak – Uniwersytet Łódzki, Wydział Nauk o Wychowaniu
Instytut Psychologii, 91-433 Łódź, ul. Smugowa 10/12

RECENZENT
Grażyna Wieczorkowska-Wierzbińska

REDAKTOR INICJUJĄCY
Urszula Dzieciątkowska

REDAKTOR WYDAWNICTWA UŁ
Katarzyna Gorzkowska

SKŁAD I ŁAMANIE
AGENT PR

PROJEKT OKŁADKI
Katarzyna Turkowska

Zdjęcie wykorzystane na okładce: © Depositphotos.com/benjaminet

Wydane przez Wydawnictwo Uniwersytetu Łódzkiego
Wydanie I. W.08579.18.0.M

Ark. wyd. 9,4; ark. druk. 13,375

ISBN 978-83-8142-211-6
e-ISBN 978-83-8142-212-3

https://doi.org/10.18778/8142-211-6

Wydawnictwo Uniwersytetu Łódzkiego
90-131 Łódź, ul. Lindleya 8
www.wydawnictwo.uni.lodz.pl
e-mail: ksiegarnia@uni.lodz.pl
tel. (42) 665 58 63

Spis treści

Przedmowa

Książkę tę napisałem w celu przypomnienia użytkownikom metod statystycznych ograniczeń, jakie one mają. Są to ograniczenia wynikające bezpośrednio z aksjomatyki, np. teorii testowania hipotez statystycznych, a także ograniczenia będące konsekwencją niedoskonałości używanego oprogramowania statystycznego. Podobnie jak nie istnieje prawdziwy model, tak też nie istnieje program komputerowy całkowicie wolny od błędów. Jednak wskazywanie błędów w oprogramowaniu nie było moim celem.

Publikacja ta nie jest podręcznikiem metod statystycznych. Ma ona na celu zwrócenie uwagi badaczom opracowującym wyniki swoich badań m.in. metodami statystycznymi, że u podstaw każdej stosowanej metody statystycznej leżą założenia, które umożliwiają sformułowanie i udowodnienie pewnych twierdzeń. Twierdzenia te z kolei pozwalają określić właściwości, np. testów statystycznych. Jeśli założenia twierdzenia nie będą spełnione, to samo twierdzenie przestaje być prawdziwe. Konsekwencją rozbieżności między teorią a praktyką statystyki są próby określenia sposobów umożliwiających stosowanie metod statystycznych mimo niespełniania przez materiał empiryczny założeń teoretycznych (np. odporność metod, mierniki oceny wielkości efektu i mnóstwo innych – mniej lub bardziej udanych – pomysłów).

Innym poważnym zagrożeniem poprawności stosowanych metod statystycznych jest dostępność oprogramowania statystycznego. Oczywiste jest, że komputer to tylko maszyna, która policzy prawie wszystko, ale tylko policzy. Używanie zaawansowanego oprogramowania przez osoby nieznające podstaw statystyki często będzie prowadziło do podejmowania irracjonalnych decyzji. Dlatego też zwracam w tej książce uwagę na tzw. kult istotności statystycznej, któremu pod żadnym pozorem nie wolno nam ulegać. Jest on bardzo wygodny, ponieważ – mówiąc brutalnie – zwalnia badacza z myślenia. A przecież to ostatnia rzecz, z której powinniśmy się zwalniać.

Przedstawione w książce zagadnienia nie wyczerpują wszystkich problemów związanych ze stosowaniem metod statystycznych. Należą do nich m.in. błędy w nazewnictwie wynikające z niepoprawnego tłumaczenia terminologii statystycznej czy błędne nazwy pewnych wyników. Sporym problemem są też

różnice w zaimplementowanych szczegółowych rozwiązaniach niektórych metod w różnych programach statystycznych.

Będę szczęśliwy, jeśli choć kilku osobom ułatwię bardziej świadome korzystanie z oprogramowania i metod statystycznych.

Autor

Wstęp

Jedno ze znaczeń słowa „statystyka" brzmi: „nazwa dyscypliny naukowej, będącej gałęzią matematyki i posiadającej własny zestaw narzędzi i metod". Chodzi tu o statystykę matematyczną. Będę ją nazywał statystyką teoretyczną albo teorią statystyki, dla podkreślenia jej dedukcyjnego charakteru. Natomiast stosowanie, wykorzystywanie metod statystycznych w praktyce – to proces indukcyjny. I ta dwoistość – dedukcja w teorii i indukcja w wykorzystaniu – prowadzi do bardzo poważnych komplikacji, które będę chciał zasygnalizować, a niektóre, być może, próbować wyjaśnić.

Jeszcze zdanie usprawiedliwienia dla tytułu książki. Zdecydowałem się na „praktykę wnioskowania statystycznego", gdyż główny akcent położyłem na opis istniejącego stanu stosowania metod statystycznych podczas opracowywania wyników badań ilościowych, głównie w naukach społecznych. Statystyczna analiza danych stosunkowo często wykonywana jest – z punktu widzenia teorii statystyki – nieprawidłowo. Nieprawidłowo w tym sensie, że badacz rzadko sprawdza założenia teoretyczne leżące u podstaw konkretnej metody. Ponadto, ponieważ teoria testowania hipotez statystycznych nie jest pozbawiona wad, powstają różne dziwne protezy mające eliminować te wady, lecz z kolei nie mają one podstaw teoretycznych, co często czyni ich stosowanie „działaniem magicznym".

> Opinie statystyków o statystyce są zwykle skrajnie optymistyczne, a nadużywanie statystyki stanowi *signum temporis* ubiegłych dekad. Jednak euforia wydaje się powoli przemijać i – nie tylko w statystyce – przychodzą chyba czasy większego poczucia rzeczywistości i odpowiedzialności zarazem. Statystyka powinna wracać pomału do punktu wyjścia: formalizowania i analizy wyników badań zjawisk empirycznych. Wymaga to przewartościowania wielu tradycji i odrzucenia wielu mitów.
>
> […]
>
> Mamy nadzieję, że przyczynimy się do tego naszą książką. Chcemy przedstawić ostry konflikt między tym, co możliwe a tym, co potrzebne. Chcemy też przedstawić nikłość powiązań między teorią a praktyką wnioskowania statystycznego, które nieraz ograniczają się do wzajemnych inspiracji.

> […] związki między problemem praktycznym a jego formalnym przedstawieniem są na ogół słabe, a rozwiązanie formalnego problemu może nie mieć racjonalnej interpretacji praktycznej.
>
> […] dochodzenie ojcostwa jest formalnie szczególnym problemem dyskryminacji, ale w praktycznym rozwiązywaniu tego problemu prawie nie korzysta się z teorii.

Powyższe cytaty pochodzą z przedmowy do książki pod redakcją Bromka i Pleszczyńskiej (1988) i przedstawiają realistyczną diagnozę problemów pojawiających się podczas stosowania metod statystyki matematycznej. Jednak diagnoza, iż „przychodzą chyba czasy większego poczucia rzeczywistości […]" wydaje mi się zbyt optymistyczna, Sądzę, że od czasu wydania książki Bromka i Pleszczynskiej bardzo mało zmieniło się w odbiorze i wykorzystaniu statystyki w naukach społecznych. W dalszej części mojej książki wielokrotnie będę wracał do tych problemów, aby skłonić Czytelnika do możliwie precyzyjnych przemyśleń – bardziej w kategoriach merytorycznych niż statystycznych – badanego i rozwiązywanego zagadnienia.

Należy nieco doprecyzować pierwsze zdanie z powyższych cytatów. Otóż, musimy cały czas rozróżniać statystyków teoretyków i statystyków praktyków, a także statystykę teoretyczną jako dział matematyki z wnioskowaniem dedukcyjnym jako narzędziem i praktykę statystyki, stosowanie metod statystycznych do rozwiązywania konkretnych zagadnień badawczych z wnioskowaniem indukcyjnym jako narzędziem. W sformułowaniu „opinie statystyków o statystyce są zwykle skrajnie optymistyczne" brak informacji, o których statystykach mówimy. W ogólności zdanie to nie jest prawdziwe.

Poświęćmy chwilę na przyjrzenie się skutkom przyjmowanej aksjomatyki, w tym momencie niezwiązanej ze statystyką. Zakładam, że każdy student – i nie tylko student – zetknął się z geometrią Euklidesa. Cały gmach tej geometrii został zbudowany około 300 r. p.n.e. na bazie pięciu aksjomatów (nazywanych także postulatami, pewnikami):

1. Od każdego punktu można poprowadzić prostą do każdego innego punktu.
2. Odcinek można dowolnie przedłużyć do linii prostej.
3. Z dowolnego środka można opisać okrąg o dowolnym promieniu.
4. Wszystkie kąty proste są równe.
5. Jeśli prosta, przecinająca dwie inne proste, tworzy z nimi po jednej stronie kąty wewnętrzne, których suma jest mniejsza od dwóch kątów prostych, to obie te proste, przedłużone nieograniczenie, przetną się po tej stronie, gdzie leżą kąty o sumie mniejszej od dwóch kątów prostych.

Aksjomat piąty w późniejszym okresie był formułowany na wiele różnych sposobów. John Pleyfair ujął go w wyjątkowo prostej formie: „Przez punkt na zewnątrz danej linii prostej może przechodzić tylko jedna prosta równoległa do niej" (Gomez, 2012).

Łobaczewski i Bolyai, niezależnie od siebie, w XIX w. zmienili aksjomat piąty: „przez punkt *P* leżący poza daną prostą *l* przechodzi więcej niż jedna prosta równoległa do prostej *l*". I zestaw aksjomatów Euklidesa ze zmienionym aksjomatem piątym stanowi podstawę geometrii hiperbolicznej. Powstała nowa geometria.

Riemann, także w wieku XIX, napisał: „Euklides twierdzi, że przez punkt poza prostą przechodzi dokładnie jedna prosta równoległa do danej prostej; Łobaczewski uważa, że takich prostych jest nieskończenie wiele; ja natomiast jestem przekonany, że takich prostych nie ma wcale". I ten zbiór aksjomatów jest podstawą jeszcze innej geometrii, geometrii eliptycznej (na sferze).

Wyniki badań Łobaczewskiego i Riemanna zostały wykorzystane przez Einsteina w definicji czasoprzestrzeni i teorii względności (Gomez, 2012).

Każda z tych geometrii jest tak samo sensowna, każda z nich jest prawdziwa, z tym że każda znajduje zastosowanie w innej sytuacji. Można powiedzieć, że w innej „rzeczywistości".

Na gruncie teorii prawdopodobieństwa i statystyki też mamy do czynienia z aksjomatyką pewnych pojęć. Na przykład pojęcie prawdopodobieństwa, podstawowe we wnioskowaniu statystycznym, zbudowane jest na bazie trzech aksjomatów, sformułowanych w 1933 r. przez Kołmogorowa. Ta aksjomatyka wyparła wcześniejsze, jako że jest bardziej nośna, lepiej nadaje się do budowy całego gmachu teorii probabilistycznej, której elementy znajdują zastosowanie w statystyce teoretycznej. Sformułowanie tych aksjomatów podaję za Fiszem (1969):

Pewnik 1. Każdemu zdarzeniu losowemu A odpowiada określona liczba $P(A)$, zwana prawdopodobieństwem zdarzenia A, spełniająca nierówność:

$$0 \leq P(A) \leq 1 \tag{0.1}$$

Pewnik 2. Prawdopodobieństwo zdarzenia pewnego równa się jedności:

$$P(\Omega) = 1 \tag{0.2}$$

Przez Ω oznaczamy przestrzeń zdarzeń elementarnych, czyli zbiór wszystkich możliwych wyników doświadczenia losowego.

Pewnik 3. Prawdopodobieństwo alternatywy skończonej lub przeliczalnej ilości zdarzeń losowych parami wyłączających się (rozłącznych) równa się sumie prawdopodobieństw tych zdarzeń.

Rozziew między teorią statystyki i zastosowaniami statystyki wyraźnie widać podczas lektury podręczników „do statystyki". O ile podręczniki pisane przez statystyków prezentują konkretne, możliwe do udowodnienia, zagadnienia – niestety, formułowane w języku matematyki – o tyle podręczniki pisane np. przez

psychologów, pedagogów i ogólniej badaczy w naukach społecznych bardzo często w ogóle nie liczą się z uwarunkowaniami teoretycznymi. Przykładowo, nieważne, jak duże prawdopodobieństwo uzyskano w teście, gdyż i tak będzie szacowana „wielkość efektu" i wykorzysta się taki wynik. Podejście takie uważam za nadmierny pragmatyzm, niekiedy za fałszowanie rezultatów badania. Opatrzenie cudzysłowem „wielkości efektu" jest wyrazem niepokoju, jaki wywołuje we mnie ten konstrukt. Będę do niego wielokrotnie wracał.

Poniżej zamieszczam konkretny przykład wnioskowania dedukcyjnego (teoria statystyki) – twierdzenie dotyczące statystyki *t*-Studenta wykorzystywanej przy porównywaniu dwóch wartości oczekiwanych (test *t*-Studenta dla prób niezależnych).

„Jeśli $\xi_1, \ldots, \xi_n, \eta_1, \ldots, \eta_k$ są niezależnymi zmiennymi losowymi o tym samym rozkładzie normalnym z wartością oczekiwaną μ i odchyleniem standardowym σ, to statystyka:

$$u = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{k+n}{kn} \cdot \frac{ns_1^2 + ks_2^2}{k+n-2}}} \tag{0.3}$$

gdzie:

$$ns_1^2 = \sum_{i=1}^{n} \left(\xi_i - \bar{x}\right)^2, \ \bar{x} = \frac{1}{n}\left(\xi_1 + \cdots + \xi_n\right) \tag{0.4}$$

$$ks_2^2 = \sum_{j=1}^{k} (\eta_j - \bar{y})^2, \ \bar{y} = \frac{1}{k}(\eta_1 + \cdots + \eta_k) \tag{0.5}$$

ma rozkład *t*-Studenta z $k + n - 2$ stopniami swobody" (Zubrzycki, 1970).

Przytoczone twierdzenie daje się udowodnić, ale czy znajduje ono zastosowanie w praktyce wykorzystywania metod statystycznych? Jest stosunkowo trudno w bezpośredni sposób wykorzystywać zarówno to, jak i inne twierdzenia statystyki teoretycznej, gdyż w praktyce niezwykle rzadko materiał empiryczny spełnia założenia twierdzenia. A jeśli założenia nie są spełnione, to twierdzenie przestaje być prawdziwym.

W tej publikacji chciałbym przedstawić Czytelnikowi pewne aspekty wnioskowania statystycznego, aksjomatyczne podstawy tego wnioskowania oraz będące konsekwencją różnej aksjomatyki problemy pojawiające się podczas stosowania metod statystycznych w praktyce. Innego typu konsekwencją problemów wnioskowania statystycznego, wynikających z różnych aksjomatyk, jest negowanie przydatności testowania hipotez przez wielu badaczy i poszukiwanie innych rozwiązań, np. ocenianie wielkości efektu, które też nie jest odbierane przez wszystkich jednoznacznie pozytywnie. Te zagadnienia również spróbuję naświetlić w jednym z rozdziałów.

Oczywiście, najpierw należy sprecyzować pojęcie wnioskowania statystycznego, co już samo w sobie jest zadaniem dość skomplikowanym. W tym obszarze

funkcjonują dwa przenikające się pojęcia. Istnieje pojęcie wnioskowania statystycznego (*statistical inference*) i pojęcie dowodu statystycznego (*statistical evidence*).

Pojęcia wnioskowania statystycznego zaczęto używać niedługo po sformułowaniu aksjomatyk testowania hipotez statystycznych, zatem mówiąc o wnioskowaniu statystycznym, będziemy traktowali je jako podejmowanie decyzji wykorzystujących metody testowania hipotez statystycznych i estymacji, zarówno punktowej, jak i przedziałowej.

Inna ważna kwestia związana z poprawnością wnioskowania statystycznego to zagadnienie mocy zastosowanego testu. Praktycznie wszystkie wykorzystywane przez nas testy statystyczne nie kontrolują prawdopodobieństwa błędu drugiego rodzaju, co znakomicie utrudnia nam poprawne wnioskowanie na podstawie wyników tych testów. Zagadnienie mocy testu i szacowania mocy testu na podstawie próby również będzie przedmiotem jednego z rozdziałów. Oczywiście, moc testu jest nierozerwalnie związana z wielkością próby, więc nie uciekniemy od kwestii szacowania wielkości próby w konkretnych analizach.

Zagadnienia poruszane w tej książce w żadnej mierze nie wyczerpują wszystkich aspektów wnioskowania statystycznego. Moim celem było zwrócenie uwagi na pewne problemy związane ze stosowaniem metod statystycznych i w konsekwencji – skłonienie badacza do refleksji nad wnioskami sformułowanymi na podstawie przeprowadzonego badania i analizy statystycznej uzyskanych danych. Chciałbym ograniczyć do minimum sytuacje, w których koronnym argumentem badacza jest: „tak wyszło z komputera" albo „tak wyszło z obliczeń". Bez względu na to, co „wyjdzie" z obliczeń, nic nie zwalnia badacza z konieczności, obowiązku – podkreślam – konieczności analizy uzyskanych wyników w terminach merytorycznych.

Praktycznie wszystkie przykłady są moimi obliczeniami w odpowiednich pakietach statystycznych. Wykorzystywane w nich były wyniki dwóch badań: wykonanych przez Bohdana Dudka, który oceniał wpływ stresu związanego z pracą na stan zdrowia pracowników służb mundurowych (Dudek, 2007) oraz przeprowadzonych przez Agnieszkę Kubot, w którym porównywano efektywność trzech rodzajów terapii w leczeniu „łokcia tenisisty" (Kubot, 2017). W przykładach wykorzystujących inne zbiory danych odpowiednie informacje podawałem w tych przykładach.

Zamieszczane w moim tekście tłumaczenia angielskojęzycznych cytatów są stosunkowo dosłowne i najczęściej zawierają nieuzasadnione (w oryginale, a więc i w tłumaczeniu) „skróty myślowe", by jak najdokładniej oddać zawarte w nich nieprawidłowości. Dlatego też, dla uniknięcia zarzutu złośliwości przy tłumaczeniu, w przypisach podaję brzmienie oryginalne.

Jeszcze drobna uwaga redakcyjna. W przytaczanych dalej przykładach będą pojawiały się oryginalne wydruki z pakietów statystycznych: SPSS 24, STATA 13 i SYSTAT 13. Starałem się możliwie niewiele ingerować w zawartość tych wydruków, choć często zawierają one zbyt dużo informacji. Zależy mi, aby Czytelnik miał możliwie pełny obraz prezentowanego zagadnienia.

Rozdział 1. Wnioskowanie statystyczne

1.1. Wprowadzenie

Opracowując wyniki badań z wykorzystaniem statystycznych metod analizy, podejmujemy decyzje będące konsekwencjami zastosowanych metod pomiaru czy klasyfikacji. Decyzje te dotyczą zarówno wyników testowania hipotez statystycznych, jak i zagadnień estymacji, głównie parametrów. Jednak kontrowersje, które budzą podejmowane decyzje odnoszą się w zasadzie wyłącznie do zagadnień testowania hipotez statystycznych. Kontrowersje te sięgają znacznie głębiej niż tylko decyzje podejmowane w wyniku testowania, gdyż wielu użytkowników metod statystycznych stawia pod znakiem zapytania także podstawy, czyli samą aksjomatykę teorii testowania hipotez statystycznych. Lambdin (2012) sugeruje w tytule swojego artykułu, że testowanie hipotez to „czarnoksięstwo". Opinię swoją opiera na fakcie ulegania kultowi istotności statystycznej przez badaczy w naukach społecznych. Niestety, w podsumowaniu artykułu brakuje propozycji innego podejścia do wnioskowania statystycznego, a jak sam autor przyznaje, jest to jedynie „wołanie o reformę" (*a call for reform*). Armstrong (2007) uważa, że testy istotności niszczą postęp w prognozowaniu. Twierdzi on, że statystyczne testy istotności są szkodliwe dla rozwoju wiedzy naukowej, ponieważ odwracają uwagę badacza od użycia **odpowiednich metod**[1]. Z kolei Loftus (1996) twierdzi, że psychologia byłaby znacznie lepszą nauką, gdyby zmienić sposób analizy danych. Zmiana sposobu analizy miałaby, według autora, polegać np. na prezentowaniu danych w postaci wykresów zamiast w tabelach z odpowiednimi prawdopodobieństwami, podawaniu przedziałów ufności (co nie jest żadnym odkryciem, chociaż nie dla wszystkich parametrów ma sens), realizowaniu metaanaliz i używanie wielkości efektu. Warto tu zwrócić uwagę, że metaanaliza jest czymś innym niż analiza danych konkretnego badania – jest to analiza opublikowanych wyników wielu badań dotyczących tego samego zagadnienia.

1 „[…] test of statistical significance are harmful to the development of scientific knowledge because they distract the researcher from the use of **proper methods**" (pogr. moje – W.Sz.).

Istnieją także opinie przeciwne, np. Häggström (2012) próbuje wyjaśnić, „dlaczego nauki empiryczne tak desperacko potrzebują statystyki?". Autor sądzi, że stosowanie metod statystycznych w dyscyplinach empirycznych jest nieuniknione i często bywa nadużywane. W konsekwencji dyskusja metodologiczna w tych dyscyplinach będzie często pociągała za sobą kwestie statystyczne, ale takie dyskusje bez udziału statystyków niosą ryzyko, że będą one po prostu nieprofesjonalne, nie będą dostarczały potrzebnych informacji. Z drugiej strony, Häggström przestrzega również przed uleganiem „kultowi istotności statystycznej". „Kult istotności statystycznej" rozumiany jest jako uznanie za ważniejsze, iż prawdopodobieństwo w teście jest mniejsze od przyjmowanego poziomu istotności, niż ewentualna ocena merytoryczna otrzymanej zależności.

Z kolei zagadnienia estymacji nie budzą raczej żadnych emocji i bardzo często opracowane wyniki analizy wyglądają tak, jakby zagadnienie estymacji nie istniało, co jest oczywistą nieprawdą. Na przykład, wykorzystując oszacowane wartości współczynników regresji, budujemy odpowiednią funkcję opisującą zależność między badanymi zmiennymi.

Wśród badaczy stosujących podczas opracowywania wyników swoich badań metody testowania hipotez statystycznych stosunkowo często można spotkać następującą opinię: im mniejsze prawdopodobieństwo w teście, tym istotniejszy wynik (silniejsza zależność). Przykładowo, Chmura-Kraemer i Kupfer (2006) piszą: „Typowy sposób przedstawiania wyników testowania hipotez statystycznych określa rezultat jako »statystycznie istotny« przy $p < 0{,}05$, co oznacza, że dane wskazują, iż dzieje się coś nielosowego. Gdy $p < 0{,}01$ dowody są bardziej przekonujące, a $p < 10^{-6}$ naprawdę bardzo przekonujące. Jednakże, chociaż wartość p pozwala określić, jak przekonująco dane świadczą przeciwko hipotezie zerowej o losowości, konkluzja zawsze przybiera formę: »zdarzyło się coś nielosowego«"[2]. Na ile nieprawdziwe jest to stwierdzenie i z czego ono wynika?

Problem polega głównie na tym, że wszystkie stosowane w praktyce 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 nam 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.

2 „As statistical hypothesis testing is typically performed, a 'statistically significant' result with $p < .05$ means that the data indicate that something nonrandom is going on. When $p < .01$, the evidence is more convincing, and $p < 10^{-6}$ very convincing indeed. However, the p value is a comment on how convincing the data are against the null hypothesis of randomness; the conclusion is always 'something nonrandom is going on' ".

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, lecz badacza zazwyczaj zaczyna wówczas interesować, jak silne są to relacje (słowo „wpływ" rezerwuję dla relacji przyczynowo-skutkowych, a nie statystycznych).

Interpretację wielkości prawdopodobieństwa uzyskanego w teście według Chmury-Kreaemer i Kupfera (2006) można uczynić bardziej intuicyjną. Mianowicie, prawdopodobieństwo mniejsze od 0,05 będzie oznaczało, że w jednym doświadczeniu zaszło mało prawdopodobne zdarzenie, co podaje w wątpliwość prawdziwość hipotezy zerowej. Prawdopodobieństwo rzędu np. 10^{-6} oznacza, iż w pojedynczym doświadczeniu zaszło zdarzenie „prawie niemożliwe", co tym bardziej świadczy przeciwko hipotezie zerowej.

Dość powszechna interpretacja, że im mniejsze prawdopodobieństwo uzyskane w teście, tym silniejsza zależność (tutaj w terminach merytorycznych) nie ma żadnego uzasadnienia statystycznego, choć badaczy ciągle dręczy pytanie: „a jak silna jest to 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. Osobiście skłaniałbym się do opinii, że są to dwa różne, choć powiązane ze sobą, sposoby wnioskowania. Nieco więcej informacji, które uzasadniałyby moją opinię znajdzie Czytelnik w podrozdziale 4.4 (*Merytoryczne znaczenie obserwowanych różnic i wielkość efektu*).

W następnych podrozdziałach spróbuję wskazać przyczyny obecnych problemów z interpretacją wyników testowania hipotez statystycznych oraz rzeczywiste niedoskonałości istniejących rozwiązań. Informacje zawarte w bieżącym rozdziale pozwolą Czytelnikowi uświadomić sobie, dlaczego pojawiło się coś takiego, jak pojęcie wielkości efektu, natomiast pominięcie tego rozdziału nie przeszkodzi mu w wykorzystywaniu mierników wielkości efektu.

1.2. Wnioskowanie statystyczne (*statistical inference*) i dowód statystyczny (*statistical evidence, statistical proof*)

W piśmiennictwie, szczególnie angielskojęzycznym, możemy spotkać się z dwoma pojęciami opisującymi efekt statystycznej analizy danych – są to wnioskowanie statystyczne i dowód statystyczny. I tu pojawia się pytanie, jak rozumieć pojęcie wnioskowania statystycznego, a jak dowodu statystycznego? Czy są to różne pojęcia? Jeśli tak, to czym się różnią? Jeśli nie, to po co istnieją obydwa?

Należy zwrócić uwagę, że mówimy o wnioskowaniu statystycznym, czyli wnioskowaniu wykorzystującym metody statystyczne. Zatem wnioski, na tym etapie rozważań, są formułowane w terminach statystycznych. Young i Smith (2005) proponują następującą definicję wnioskowania statystycznego: „we wnioskowaniu statystycznym dane pochodzące z eksperymentu albo badania obserwacyjnego są modelowane jako obserwowane wartości zmiennych losowych, by dostarczyć pewnych ram umożliwiających wyciąganie indukcyjnych wniosków o mechanizmach działających w populacyjnych danych".

Proponowałbym pojęcie nieco bardziej intuicyjne: przez wnioskowanie statystyczne będziemy rozumieli postępowanie wykorzystujące metody statystyczne, umożliwiające uogólnienie zależności zaobserwowanych w próbie na populację generalną, z której ta próba pochodzi.

Pojęcie próby oraz sposoby jej doboru zostały krótko omówione w podręczniku Szymczaka (2018). Więcej szczegółów Czytelnik znajdzie w monografii Levy'ego i Lemeshowa (1991).

Tak rozumiane wnioskowanie statystyczne jest na tyle ogólne, że praktycznie nie zostało przypisane do żadnego z paradygmatów statystycznych.

A co z dowodem statystycznym? Bill Thompson (2007) próbuje nakreślić pewne ramy dla pojęcia dowodu statystycznego, przyznając jednak, że nie potrafimy precyzyjnie go zdefiniować. Thompson wiąże pojęcie dowodu statystycznego z eksperymentem, zauważając przy tym, iż pojęcia eksperymentu, parametru i dowodu odgrywają centralną rolę w teorii statystyki, a mimo to ich znaczenie jest często starannie pomijane. Zatem sądzę, że pojęcie dowodu statystycznego jest innym zwrotem – służącym uniknięciu powtórzeń – na określenie wnioskowania statystycznego. I zamiast korzystać z jakiegoś automatycznego miernika siły dowodu statystycznego, badacz będzie musiał przeprowadzić merytoryczną interpretację wyników analizy statystycznej i podjąć odpowiednią decyzję, już w terminach merytorycznych.

1.3. Paradygmaty statystyki

Pojęcie paradygmatu zostało zaproponowane dopiero w 1962 r. przez Thomasa S. Kuhna (Kuhn, 2009). W czasach „przedparadygmatowych" występowało pojęcie podstaw statystyki czy podstaw wnioskowania statystycznego. Poniżej przedstawię dyskusję, jaka toczyła się na temat podstaw statystyki (*foundations of statistics*) w latach 1958–1962 po ukazaniu się książki Savage'a w 1954 r. (Savage, 1954). Ale także w późniejszym okresie zdarzały się artykuły, których autorzy, dyskutując o podstawach statystyki, nie używali pojęcia paradygmatu, np. Efron (1978) czy Freedman (1995/1996).

Autorzy analizujący podstawy statystyki na ogół zgadzają się, że podstawy statystyki są kontrowersyjne i zmienne. Jednakże są one częścią podstaw nauki w najszerszym sensie i mimo niezbędności wykorzystywania w statystyce narzędzi matematycznych, jej podstawy mogą być rozważane w aspekcie filozoficznym (Savage, 1958). Z innego punktu widzenia, statystyka wydaje się trudna również dla matematyków – być może z powodu nieosiągalności tradycyjnej metody przedstawiania wyników w matematyce poprzez twierdzenie-dowód. Słabym pocieszeniem wydaje się fakt, iż statystyka jest trudna i dla statystyków (Efron, 1978).

Toczącej się od kilkudziesięciu lat (niektórzy uważają, że od ponad dwustu lat) dyskusji na temat podstaw statystyki nie widać końca. W chwili obecnej wcale nie jesteśmy bliżsi ujednolicenia podejścia do metod statystycznych niż byliśmy kilkadziesiąt lat wcześniej. Nadal toczą się spory między wyznawcami podejścia częstotliwościowego i wyznawcami podejścia bayesowskiego. Nieco inaczej granica ta przebiega między wyznawcami obiektywności i subiektywności w statystyce. W grupie zwolenników podejścia częstotliwościowego też występują różnice między opowiadającymi się za „wnioskowaniem indukcyjnym" według teorii Fishera i „postępowaniem indukcyjnym" według teorii Neymana-Pearsona. Dyskusje te toczą się wśród matematyków i statystyków. Spróbujmy wyobrazić sobie, co dzieje się wśród badaczy wykorzystujących metody statystyczne do opracowywania wyników swoich badań, którzy nie mają wiedzy matematycznej. W rozdziałach trzecim i czwartym przedstawię pewne ich propozycje rozwiązania problemu.

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 jakaś 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 jednak 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 tym stwierdzeniem pokazuje, na czym polega problem z prawdopodobieństwem. Bo cóż oznacza duży stopień pewności? Jest to po prostu inna nazwa stosunkowo dużego prawdopodobieństwa. Zatem cytowane zdanie nic nie wyjaśnia. I trzeba 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). Czy jednak należy zgodzić się ze stwierdzeniem, że statystyka jest dyscypliną rozwiniętą z jasno sformułowanych aksjomatów? Raczej różne statystyki są rozwijane z jasno formułowanych różnych zbiorów aksjomatów.

Ale wróćmy do zagadnień prawdopodobieństwa zdarzenia. Brak intuicji prawdopodobieństwa zdarzenia spowodował powstanie kilku definicji prawdopodobieństwa, co doskonale utrudnia późniejsze wykorzystywanie tego pojęcia w analizach statystycznych. Możemy wyróżnić:

- definicję aksjomatyczną (Kołmogorow),
- definicję klasyczną (Laplace),
- definicję wykorzystującą częstości względne (von Mises),
- prawdopodobieństwo jako miarę przekonania.

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:

$$P(\bigcup_k A_k) = \sum_k P(A_k);\ \ A_i \cap A_j = \emptyset\ \ i,j = 1,2,\cdots,k;\ i \neq j \tag{1.1}$$

Wzór ten można zapisać w nieco innej postaci:

$$\begin{aligned} &P(\mathrm{A}_1 \cup A_2 \cup \cdots \cup A_k \cup \cdots) = \\ &= P(A_1) + P(A_2) + \cdots + P(A_k) + \cdots;\ A_i \cap A_j = \emptyset\ \ i,j = 1,\cdots,k;\ i \neq j \end{aligned} \tag{1.2}$$

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

$$P(\Omega) = 1 \tag{1.3}$$

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

$$P(\emptyset) = 0. \tag{1.4}$$

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.

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:

$$P(A) = \frac{k}{n} \tag{1.5}$$

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". Wyobraźmy sobie eksperyment polegający na rzucie regularną (prawidłową) sześcienną kostką do gry i rozważmy zdarzenie polegające na wyrzuceniu nieparzystej liczby oczek. 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 $\frac{k}{n}$ nie daje się obliczyć nawet wtedy, gdy k jest liczbą skończoną. Wówczas zamiast liczby elementów musimy używać innych liczb, zwanych mocami zbiorów, pełniących podobną rolę jak liczebności, lecz będzie to już inna definicja.

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:

$$P(A) = \lim_{n\to\infty} \frac{n_A}{n} \tag{1.6}$$

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 oscylowały wokół pewnej stałej wartości, wokół liczby będącej prawdopodobieństwem $P(A)$. Ale, 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 także zastosowanie, gdy przestrzeń zdarzeń elementarnych zawiera nieskończoną ilość elementów.

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 mylnego – jest ono przedmiotem wielu artykułów, nie tylko z zakresu zastosowań, lecz także teorii, zob. np. Anscombe i Aumann (1963), w którym znalazł się rozdział o istnieniu subiektywnych prawdopodobieństw, Machina i Schmeidler (1992) czy Karni (1993).

1.3.1. Paradygmat w nauce

Nim przejdę do rozważania paradygmatów statystycznych, chciałbym przybliżyć Czytelnikowi pojęcie paradygmatu w nauce, zaproponowane przez Thomasa Samuela Kuhna (Kuhn, 2009). Na pewnym etapie swoich badań z historii nauki Kuhn skonstatował, że rozwój nauki nie odbywa się w sposób kumulatywny poprzez gromadzenie coraz większej ilości informacji, lecz istotniejsze dla tego rozwoju są, występujące co pewien czas, rewolucje naukowe. Zgromadzone na pewnym etapie rozwoju wiedzy informacje przestają być zgodne z dotychczasowymi przesłankami (według Kuhna paradygmatami) i zostają zastąpione innymi.

„Termin »nauka normalna« oznacza w niniejszych rozważaniach badania wyrastające z jednego lub wielu takich osiągnięć przeszłości, które dana społeczność uczonych aktualnie akceptuje i traktuje jako fundament swej dalszej praktyki”[3]. I najważniejsze – pojęcie paradygmatu, choć, jak przyznaje sam Kuhn w napisanym w 1969 r. *Postscriptum*, „nie ma bardziej niejasnej i ważniejszej kwestii w moim pierwotnym tekście. Pewna życzliwa czytelniczka podzielająca moje przeświadczenie, że w pojęciu paradygmatu skupiają się najważniejsze filozoficzne treści książki, przygotowała częściowy indeks analityczny i doszła do wniosku, że terminu tego używa się w niej na co najmniej dwadzieścia dwa różne sposoby”.

Otóż, paradygmatami Kuhn będzie nazywał (to najprawdopodobniej jeden z dwudziestu dwóch sposobów rozumienia tego pojęcia) osiągnięcia naukowe

3 Wszystkie cytaty w tym rozdziale pochodzą z wersji książki Kuhna z 2009 r.

oznaczające się na tyle dużą oryginalnością i atrakcyjnością, „aby odwrócić uwagę stałej grupy zwolenników danej teorii od konkurencyjnych sposobów uprawiania nauki. Jednocześnie dorobek ten ma być na tyle otwarty, żeby pozostawiał nowej szkole najrozmaitsze problemy do rozwiązania". „Pojęcie paradygmatu pozostaje w ścisłym związku z pojęciem nauki normalnej".

Ale w *Postscriptum* Kuhn stwierdza, że: „Paradygmat rządzi w pierwszej kolejności nie dziedziną przedmiotową, lecz raczej grupą praktykujących uczonych. Wszelka analiza badań kierujących się paradygmatem bądź też rozbijających go musi zacząć od zlokalizowania odpowiedniej grupy czy grup". Nastąpiło więc wyraźne przeniesienie pojęcia paradygmatu z osiągnięć naukowych na grupę badaczy. Sądzę, że bardzo zmienia to sens pojęcia paradygmatu.

1.3.2. Paradygmaty w statystyce

Pojęcia paradygmatu będę używał w dalszej części tego opracowania w sensie podkreślonym w *Postscriptum* Kuhna, czyli charakteryzującym grupę badaczy. W przypadku statystyki oraz – co ważniejsze – jej zastosowań takie rozumienie paradygmatu będzie zdecydowanie klarowniejsze. Jak zauważa Efron (1978), jego artykuł próbuje spenetrować filozoficzne boje toczone między wyznawcami podejścia bayesowskiego (*Bayesians*), statystykami klasycznymi (częstotliwościowymi) (*frequentists*) i trzecią grupą określaną jako „fisherowcy" (*Fisherians*), a więc między grupami uczonych. Oczywiście owe trzy podejścia różnią się także zbiorami aksjomatów leżących u podstaw każdej z tych teorii, ale nie mam pewności, czy z tego powodu można je nazywać paradygmatami w sensie osiągnięć naukowych.

W podręczniku pod redakcją Bandyopadhyaya i Forstera (2011) przedstawione zostały cztery paradygmaty statystyki: paradygmat klasyczny (*classical statistics paradigm*), paradygmat bayesowski (*Bayesian paradigm*), paradygmat wiarygodnościowy (*likelihood paradigm*) i paradygmat Akaike (*Akaikean paradigm*). Natomiast Aitkin (2011) uważa, że w statystyce najczęściej mówi się o paradygmacie bayesowskim i rzadziej, ale jednak, możemy też mówić o paradygmacie częstościowym. Paradygmat nazywany przez jednych autorów klasycznym – lub częstościowym przez innych – jest tym samym paradygmatem. Paradygmaty częstościowy i bayesowski są na tyle różne, że spowodowały powstanie odrębnych szkół statystycznych (grup „wyznawców"). I, jak zauważa Aitkin, dopóki te dwa podejścia są stosowane do rozwiązywania różnych problemów z wykorzystaniem metod statystycznych, wszystko jest w porządku. Natomiast pojawiają się poważne kłopoty, gdy te dwie szkoły uzurpują sobie prawo do bycia uniwersalnymi w szerokim zakresie zastosowań metod statystycznych. Aitkin (2011) wymienia ponadto paradygmat wiarygodnościowy, lecz nim się nie zajmuje.

Young i Smith (2005) w ślad za Efronem (1978) identyfikują trzy główne paradygmaty wnioskowania statystycznego: bayesowski, fisherowski (utożsamiany

przez niektórych autorów z paradygmatem wiarygodnościowym, zob. np. Efron, 1998) i częstościowy.

Pojęcia paradygmatu i wnioskowania statystycznego są pojęciami bardzo ogólnymi, nawet zbyt ogólnymi dla celów tego opracowania, dlatego w tym momencie ograniczę się do zagadnienia testowania hipotez statystycznych. Zrezygnuję też z używania pojęcia paradygmatu statystyki, stosując w zamian konkretniejsze pojęcia aksjomatyki albo teorii testowania hipotez.

Rozwiązania bayesowskie w psychologii i innych naukach społecznych oraz medycznych stosowane są bardzo rzadko. Pojawiające się w tych naukach kontrowersje interpretacyjne dotyczące uzyskanych wyników analiz danych nie są wynikiem stosowania różnych paradygmatów statystyki, lecz rezultatem pomieszania dwóch różnych aksjomatyk: aksjomatyki Fishera i aksjomatyki Neymana-Pearsona.

Rozmyślnie użyłem tutaj sformułowania „aksjomatyki", choć może przesłanki sformułowane przez ich twórców trudno nazwać aksjomatami, ale chciałem zwrócić uwagę, że każda teoria funkcjonuje w ramach pewnego zbioru „przesłanek fundamentalnych", aksjomatów, które są przyjmowane bez udowadniania, gdyż udowodnione być nie mogą. Stosowanie takich teorii sprowadza na badacza różne skutki – często takie, których sobie nie życzy. Próbuje on wówczas je wyeliminować, niekiedy w irracjonalny sposób.

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 takie, które jest zgodne z teorią 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 kilku „teorii statystyki" były przedmiotem zainteresowania matematyków i statystyków (Inman, 1994; Lehmann, 1993, 1995; Berger, 2003; Christensen, 2005; Manthey, 2010).

1.4. 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 i 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.

1.5. Teoria Neymana-Pearsona

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

$$\begin{cases} H_0: \theta = \theta_0 \\ H_1: \theta = \theta_1 \end{cases} \tag{1.7}$$

Ich sposób postępowania 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 że w wielokrotnie powtarzanych badaniach z użyciem procedury statystycznej, uzyskiwany na dłuższą metę, przeciętny błąd (częstość podjęcia błędnej decyzji) nie powinien być większy niż określone z góry prawdopodobieństwo (Neyman, 1977). Neyman i Pearson całkowicie rozwiązali problem testowania w przypadku prostej hipotezy zerowej oraz prostej hipotezy alternatywnej (lemat Neymana-Pearsona). Jednak dla bardziej złożonych przypadków, złożonych hipotez alternatywnych, teoria wymagała dodatkowych pomysłów i opracowywanie szczegółów tego programu było głównym przedmiotem zainteresowań statystyki matematycznej (teoretycznej) w następnych dekadach.

1.5.1. 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:

- hipotezę nazywamy prostą, gdy określa ona jednoznacznie rozkład prawdopodobieństwa; każdą hipotezę, która nie jest prostą, nazywamy 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, czyli mówimy o błędnych decyzjach polegających na odrzuceniu prawdziwej hipotezy zerowej.

A dlaczego nie odnosi się 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 prostą, a alternatywna złożoną? Nie musi tak być. Na przykład Rao (1965) rozważa sytuacje, w których zarówno hipoteza zerowa, jak i alternatywna są hipotezami złożonymi:

Tabela 1.1. Przypadek złożonej hipotezy zerowej i złożonej hipotezy alternatywnej

H_0	H_1
$\Theta \leq \theta_0$	$\Theta > \theta_0$
$\Theta \leq \theta_0$ lub[4] $\theta \geq \theta_1$	$\Theta_0 < \theta < \theta_1$
$\Theta_0 \leq \theta \leq \theta_1$	$\Theta < \theta_0$ lub[4] $\theta > \theta_1$

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

$$\alpha(\theta) \leq \alpha \text{ dla } \theta \in H_0 \tag{1.8}$$

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

4 Tabela ta pochodzi z polskiego tłumaczenia podręcznika Rao (1965), w którym użyty jest spójnik „lub". Zamiast „lub" poprawniejszy jest w tym przypadku łącznik „albo". W języku polskim spójniki „lub" i „albo" traktowane są jako synonimy, natomiast w logice matematycznej są to dwa różne funktory zdaniotwórcze mające inne własności.

Silvey (1978) przedstawia „matematyczną" metodę przezwyciężenia trudności wynikających z ewentualnych nieciągłości rozkładów prawdopodobieństwa. Omawiana przez niego metoda nie dotyczy przypadku ogólnego, lecz jedynie zagadnienia testowania prostej hipotezy zerowej przeciwko prostej alternatywnej. Rozważania te prowadzą do testu najmocniejszego na poziomie istotności α. Ale w sytuacji prostej hipotezy zerowej i złożonej alternatywnej test jednostajnie najmocniejszy praktycznie 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:

$$\begin{cases} H_0: \theta = \theta_0 \\ H_1: \theta \neq \theta_0 \end{cases} \tag{1.9}$$

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 1.1.

PRZYKŁAD 1.1. Zagadnienie porównywania dwóch wartości oczekiwanych dla prób niezależnych.

$$\begin{cases} H_0: \mu_1 = \mu_2 \\ H_1: \mu_1 \neq \mu_2 \end{cases} \tag{1.10}$$

Niech $(x_{11}, x_{12}, \ldots, x_{1n1})$, $(x_{21}, x_{22}, \ldots, x_{2n2})$ 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

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{n_1 \cdot s_1^2 + n_2 \cdot s_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \tag{1.11}$$

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, które zostały udowodnione w terminach matematycznych) 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 (1.11) ma rozkład *t*-Studenta i test, który ją wykorzystuje, ma odpowiednie cechy dla testowania hipotez (1.10).

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), ale znajdujących uzasadnienie w teorii.

1.5.2. Argumenty przeciw teorii Neymana-Pearsona i za nią

Teoria Neymana-Pearsona testowania hipotez statystycznych jest w zasadzie jedyną wykorzystywaną w naukach społecznych. Mowa tu o teorii w postaci, jaką zaproponowali Jerzy Neyman i Egon S. Pearson. Włączenie do zagadnienia testowania hipotezy alternatywnej zdecydowanie zwiększyło jej efektywność w porównaniu z teorią proponowaną przez R. A. Fishera. W zasadzie fakt jej istnienia jest praktycznie jedynym argumentem „za". Stosowanie metod testowania hipotez pozwoliło, w pewnym zakresie, zobiektywizować wyniki badań i umożliwić ich porównywanie. Bardzo często badacz przywiązany i związany emocjonalnie z wynikami swoich obserwacji widzi w danych zależności, które w rzeczywistości nie istnieją, a zastosowanie metod statystycznych pozwala mu to zauważyć.

Jak zwykle, jest też druga strona medalu. Gigerenzer (2004), który jest z wykształcenia psychologiem, jasno stwierdza, że „Statystyczne rytuały przeważnie eliminują w naukach społecznych statystyczne myślenie. Rytuały są niezbędne do identyfikacji

z grupami społecznymi i powinny bardziej być przedmiotem badań niż metodą naukową". A takim rytuałem w opinii wielu badaczy jest ortodoksyjne stosowanie teorii testowania hipotez statystycznych Neymana-Pearsona, tj. traktowanie wyników testowania w terminach biało-czarnych: istotne albo nieistotne statystycznie (Gigerenzer i wsp., 2004; Gigerenzer, 2004; Ziliak, McCloskey, 2009; Krämer, 2011).

Gigerenzer (2004) próbuje ustalić przyczynę stosowania rytuałów statystycznych. Stwierdza, że podręczniki i programy nauczania psychologów prawie nigdy nie zawierają szerszego instrumentarium statystycznego, w którym powinny znaleźć się takie narzędzia, jak statystyka opisowa, metody eksploracyjne Tukeya, statystyka bayesowska, teoria decyzji Neymana-Pearsona czy analiza sekwencyjna Walda. Znajomość zawartości takiego instrumentarium oczywiście wymaga myślenia statystycznego, tj. sztuki wyboru odpowiedniego narzędzia dla danego problemu badawczego. Zamiast tego w tekstach i praktyce badacze skłaniają się do – jak Gigerenzer to nazywa – rytualnego stosowania testowania hipotezy zerowej (*null ritual*). Brzmi to ładnie, lecz moje doświadczenia w nauczaniu statystyki na studiach psychologicznych pokazują, że jest to plan niemożliwy do zrealizowania – po prostu brak u studentów znajomości podstaw matematycznych umożliwiających opanowanie takiego instrumentarium. I przykre to, ale nie można nie zgodzić się z Maslowem (1966): „jest nęcącym w sytuacji, gdy jedynym narzędziem, którym dysponujemy, jest młotek, aby spróbować każdą rzecz potraktować jakby była gwoździem". Natomiast stwierdzenie Fishera (1956), iż: „[…] żaden pracownik naukowy nie może jednego, ustalonego poziomu istotności używać rok po roku i we wszystkich okolicznościach, by odrzucić hipotezę; raczej powinien, w każdym szczególnym przypadku w świetle problemu do udowodnienia i jego idei, przemyśleć zagadnienie", odnosi się już do badaczy. I mimo że zostało sformułowane kilkadziesiąt lat temu, niestety, nic nie straciło ze swojej aktualności.

1.6. Podejście bayesowskie

W rzeczywistości, jak już wspomniałem, oprócz podejścia Fishera i Neymana-Pearsona istnieje jeszcze kilka podejść do zagadnienia wnioskowania statystycznego, np. podejście Jeffreysa, które jest podejściem bayesowskim, a także podejście wiarygodnościowe (*likelihood paradigm*), u podstaw którego również leży twierdzenie Bayesa. W podejściu bayesowskim – najogólniej mówiąc, aby nie wchodzić w zbędne w tych rozważaniach szczegóły – przyjmuje się, że występujące w modelach parametry nie są stałymi, lecz zmiennymi losowymi o pewnych, znanych albo sugerowanych, rozkładach prawdopodobieństwa. Podejmowane są też próby, choć nieczęste, wykorzystania metod bayesowskich w psychologii (Trafimow, 2003; Lee, Wagenmakers, 2005). Więcej szczegółów dotyczących rozwiązań bayesowskich przedstawiam w rozdziale piątym.

1.7. 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. 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 odrzucenia hipotezy zerowej dostarczanej przez dane. Hipoteza zerowa zostaje odrzucona np. zarówno dla $t = 2$, jak i $t = 81$ przy $\alpha = 0{,}05$. Podejście Neymana-Pearsona krytykowane było także z powodu potrzeby określania hipotezy alternatywnej i, w konsekwencji, związanych z tym trudności z określeniem prawdopodobieństwa błędu drugiego rodzaju (więcej szczegółów na temat prawdopodobieństwa błędu drugiego rodzaju i związanego z nim pojęcia mocy testu statystycznego znajdzie Czytelnik w rozdziale trzecim).

Z kolei *p* w teorii Fishera było podstawą zarzutu naruszenia częstościowej zasady prawdopodobieństwa. Warto w tym miejscu przypomnieć, że praca Kołmogorowa, w której przedstawił on układ aksjomatów prawdopodobieństwa zdarzenia, ukazała się dopiero w roku 1933, więc wydaje się, że w momencie powstawania teorii testowania hipotez statystycznych nie była jeszcze powszechnie znana. Jeffreys uważał, że logika wykorzystująca wartość *p* „na ogonie" obszaru (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, stwierdzał: „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).

Niemal powszechnie utrzymuje się pogląd, że teorie te – z jednej strony Fishera, z drugiej Neymana i Pearsona – są całkiem różne. Znajduje to odzwierciedlenie w fakcie, iż często są używane odrębne określenia (mimo że czasami niekonsekwentnie) do nazwania tych dwóch podejść: testowania istotności dla Fishera i testowania hipotez dla Neymana i Pearsona. (Ponieważ oba

dotyczą testowania hipotez, więc jest wygodne ignorować terminologiczne różnice i stosować 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). Ponadto, dyskutanci podnoszą zagadnienia unifikacji podejść, a także utrzymania istniejącej różnorodności, przytaczając argumenty zarówno za, jak i przeciw unifikacji.

Jednak w większości współczesnych opracowań metod statystycznych elementy tych dwóch niekompatybilnych podejść są mieszane, co bardzo niekorzystnie odbija się na praktyce wnioskowania (np. porównywanie prawdopodobieństw i w konsekwencji porównywanie siły zależności).

W świetle istnienia dwóch, niejako konkurencyjnych, teorii testowania hipotez statystycznych naturalne wszakże 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 wsp., 2008; Roberts, Pashler, 2000; Rodgers, 2010; Sterne, 2002; Denis, 2003; Jones, Tukey, 2000; Killen, 2005; Thompson Bruce, 1994).

Thompson Bruce (1994) zauważa, że:

> zbyt mało badaczy rozumie, co testy statystyczne „robią, a czego nie robią", i w konsekwencji wyniki ich badań mogą być 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 nie dotyczy psychologów, ale dobrze jest zdawać sobie z niego sprawę.

I jeszcze jeden cytat z Thompsona Bruce'a (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 wartość p wyraźnie 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, tworząc fałszywe wrażenie jednego, spójnego modelu wnioskowania statystycznego (Hubbard, Armstrong, 2006).

Z powodów, które naszkicowałem powyżej, 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 – jak się wydaje – nie powinny być stosowane. Co prawda, sam nie jestem bez grzechu, gdyż zwrotu „[…] w sposób istotny statystycznie […]" jednak używam. Może rozważania tego rozdziału zwrócą uwagę badaczy na konieczność stosowania jednoznacznej, precyzyjniejszej terminologii statystycznej.

W moim podręczniku (Szymczak, 2018) 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, a może stopnia zaufania do podjętej decyzji) na podstawie wartości statystyki będącej podstawą testu czy na podstawie wartości prawdopodobieństwa oszacowanego w teście.

Zrozumiała więc wydaje się próba wprowadzenia jakiejś miary – miary wielkości efektu. Mówienie o wielkości efektu dla testowania hipotez według Neymana-Pearsona wydaje się nieporozumieniem, gdyż w tym przypadku na podstawie wielkości prawdopodobieństwa podejmujemy jedną z dwóch możliwych decyzji. I „małość" prawdopodobieństwa o niczym nie świadczy. Inna sytuacja ma miejsce, gdy używamy teorii Fishera. Skoro wartość p jest interpretowana jako siła dowodu przeciwko H_0, to wartość p większą od przyjętej wartości granicznej możemy (możemy?) potraktować jako siłę dowodu przeciwko H_0 z przeciwnym znakiem (ale nie mamy prawa traktować tej wartości jako siły dowodu za prawdziwością H_0). Czyli w przypadku nieodrzucenia hipotezy zerowej warto będzie mieć instrumenty oceniające siłę zależności sformułowanej w hipotezie zerowej. Używany jednak w praktyce schemat: hipoteza zerowa *vs.* hipoteza alternatywna, gdy hipoteza zerowa jest hipotezą prostą, powoduje, że $p > \alpha$ nie prowadzi do żadnej sensownej decyzji, gdyż w znakomitej większości przypadków bardziej jesteśmy

zainteresowani przyjęciem hipotezy alternatywnej. Znajduje to również odbicie w trudności opublikowania wyników badania, w którym nie mamy „statystycznie istotnych wyników". Pamiętajmy także, iż w sytuacji wykorzystywania teorii Fishera nie dysponujemy żadną hipotezą alternatywną i odrzucenie hipotezy zerowej zmusza nas do zbudowania nowego modelu merytorycznego.

1.8. „Kult istotności statystycznej"

Rozpocznę od przykładu na rzeczywistych danych. Poniżej przedstawiony jest fragment zbioru danych badania Dudka (2007) (szczegółowe informacje o występujących w tym badaniu zmiennych Czytelnik znajdzie w załączniku 1).

Tabela 1.2. Fragment wyników badania (Dudek, 2007); w tabeli przedstawiono tylko kilka zmiennych, wybranych spośród wszystkich badanych

nr_bad	subiekt	SOC	wrogosc	zaklopot	przygneb	znuze-nie	zyczliwo	napiecie	wigor	GHQ_ suma
1	137	135	12	4	19	7	23	5	21	18
2	81	147	2	0	5	6	25	3	23	13
3	121	131	14	10	17	8	17	19	19	12
4	139	156	9	5	12	9	26	8	24	26
5	80	170	11	4	16	9	25	6	22	14
6	105	175	15	7	19	5	19	11	21	15
7	99	140	12	5	17	10	16	10	15	25
8	137	152	15	9	16	14	21	18	23	22
9	80	158	8	4	12	5	21	5	21	12
10	85	117	20	7	20	13	20	11	20	25
11	93	174	12	9	9	9	22	4	20	20
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
437	136	152	16	6	22	11	18	14	20	20
438	162	135	14	7	18	10	21	4	22	32
439	227	78	42	16	51	23	9	23	9	48
440	209	78	47	19	58	27	8	29	16	57

Tabela 1.2 (cd.)

nr_bad	subiekt	SOC	wrogosc	zaklopot	przygneb	znuze-nie	zyczliwo	napiecie	wigor	GHQ_ suma
441	138	149	0	2	14	3	16	0	20	19
442	107	177	17	2	10	10	20	9	17	16
443	99	143	11	5	15	6	23	5	24	15
444	94	131	14	7	18	11	23	13	21	16

W tab. 1.3 przedstawione zostały porachowane współczynniki korelacji liniowej między „zmienną" 'nr_bad' i dziesięcioma zmiennymi uzyskanymi w badaniu. Poniżej zamieściłem oryginalny wydruk z programu SPSS. Tylko dla współczynnika korelacji między 'nr_bad' (numerem badanego) i obserwowaną zmienną 'napięcie' prawdopodobieństwo w teście hipotezy:

$$\begin{cases} H_0: \rho = 0 \\ H_1: \rho \neq 0 \end{cases} \tag{1.12}$$

jest większe od 0,05; wszystkie pozostałe prawdopodobieństwa są mniejsze od 0,05. Zatem między „zmienną" 'nr_bad' i prawie wszystkimi przedstawionymi zmiennymi mierzonymi w badaniu i pokazanymi w przykładzie występuje istotna statystycznie zależność liniowa.

Tabela 1.3. Współczynniki korelacji między sztuczną zmienną 'nr_bad' a rzeczywistymi wynikami badania

Badane zmienne		nr_bad
subiekt	Korelacja Pearsona	,133**
	Istotność (dwustronna)	**,005**
	N	444
SOC	Korelacja Pearsona	–,129**
	Istotność (dwustronna)	**,006**
	N	444
wrogosc	Korelacja Pearsona	,121*
	Istotność (dwustronna)	**,011**
	N	444

Badane zmienne		nr_bad
zaklopot	Korelacja Pearsona	,098*
	Istotność (dwustronna)	**,039**
	N	444
przygneb	Korelacja Pearsona	,140**
	Istotność (dwustronna)	**,003**
	N	444
znuzenie	Korelacja Pearsona	,124**
	Istotność (dwustronna)	**,009**
	N	444
zyczliwo	Korelacja Pearsona	–,183**
	Istotność (dwustronna)	**,000**
	N	444
napiecie	Korelacja Pearsona	,050
	Istotność (dwustronna)	,297
	N	444
wigor	Korelacja Pearsona	–,117*
	Istotność (dwustronna)	**,014**
	N	444
GHQ_suma	Korelacja Pearsona	,096*
	Istotność (dwustronna)	**,044**
	N	440

** Korelacja istotna na poziomie 0,01 (dwustronnie).
* Korelacja istotna na poziomie 0,05 (dwustronnie).

Mam nadzieję, że wyraźnie widać na podstawie obliczeń w tym przykładzie, iż „istotność statystyczna" bez oceny merytorycznej badanej zależności jest pustym pojęciem. Jest to, co prawda, przykład mocno przerysowany, ale w praktyce spotykamy się z „istotnymi zależnościami", które wydają się pozornie prawdziwe i dlatego stwarzają poważne zagrożenie.

Chciałbym zwrócić uwagę Czytelnika na tekst pod powyższą tabelką. W tabelce podane są prawdopodobieństwa w teście hipotezy (1.12). Każda osoba analizująca zapisy w tej tabeli potrafi ocenić, jakie są relacje między prawdopodobieństwem i przyjętym w danym badaniu poziomem istotności testu. Nie ma sensu komentarz,

że jeden współczynnik jest istotny na poziomie 0,05, a inny na poziomie 0,01. O czym to świadczy? Że jedna zależność jest silniejsza od drugiej? Jeśli chcemy to ocenić, przetestujmy hipotezę o równości współczynników korelacji. A jak pisałem już wcześniej, porównywanie prawdopodobieństw to nieporozumienie. Nie ma żadnych podstaw, w żadnej aksjomatyce testowania hipotez, aby coś takiego robić.

W podobnym tonie wypowiadają się Ziliak i McCloskey (2009). Istotność, zredukowana jedynie do wąskiego i statystycznego znaczenia, np. $p < 0{,}05$, ma mało wspólnego z dającym się uzasadnić wnioskowaniem naukowym, analizą błędu czy racjonalnym podejmowaniem decyzji. I jak dotąd, w powszechnym użyciu prowadzi do powstawania ogromnej ilości niekontrolowanych strat w nauce i społeczeństwie.

Krämer (2011) idzie jeszcze dalej, sugerując, że statystyczne testowanie istotności (*statistical significance testing*) jest raczej barierą niż narzędziem wspomagającym badania empiryczne w ekonomii, powinny zatem zostać całkowicie porzucone.

Cumming (2014) pisze o nowej statystyce (*The New Statistics*), na którą mają składać się: wielkość efektu, przedziały ufności i metaanalizy. Gigerenzer (2004), Gigerenzer i wsp. (2004) oraz Gigerenzer i Marewski (2015) również podnoszą problem statystycznego rytuału zabijającego statystyczne myślenie, proponując zmiany w nauczaniu statystyki, aby poszerzyć zakres metod. Jednak nikt do tej pory nie zaproponował spójnej, o solidnych podstawach matematycznych, teorii testowania hipotez na miarę teorii Fishera czy Neymana-Pearsona.

1.9. Podsumowanie

Na początek zacytuję fragment artykułu Rodgersa (2010): „Zajmowanie się nauką jako procesem twórczym wymaga naukowego twórczego myślenia. Stosowanie NHST (*Null Hypothesis Significance Testing*) [czyli po prostu metod testowania hipotez statystycznych – przyp. W.Sz.] jako technicznego zbioru procedur wyklucza kreatywność".

Ale chciałbym także zwrócić uwagę Czytelnika na, co prawda nieliczne, artykuły broniące metod NHST. Hagen (1997) dyskutuje z poglądami Cohena (1994), który według Hagena uważa, że:

- NHTS nie mówią nam tego, co chcielibyśmy wiedzieć;
- hipoteza zerowa jest zawsze fałszywa;
- NHTS brak logicznej spójności.

Główny zarzut Hagena w stosunku do interpretacji Cohena polega na tym, iż w różnych przykładach Cohena H_0 i H_1 odnoszą się do próby, a nie do populacji generalnej. A przecież wszystkie hipotezy statystyczne odnoszą się do populacji generalnej. I to powoduje, że np. nie jesteśmy w stanie określić mocy testu, gdyż zależy on od konkretnej hipotezy będącej składową złożonej hipotezy alternatywnej (więcej na temat tego zagadnienia piszę w rozdziale trzecim).

Robinson i Wainer (2001) zadają pytanie: „Czym jest efekt, o którego wielkości mówimy?". W wielu sytuacjach wielkość efektu o niczym nie świadczy, zaś istotniejszy jest praktyczny, merytoryczny efekt, niekiedy wystarczy kierunek zależności. Autorzy ci zadają jeszcze jedno ciekawe i trudne pytanie: „Co, jeśli $p = 0{,}06$?". Na tak postawione pytanie nie ma prostej, jednoznacznej odpowiedzi. Autorzy uważają, że wartości p powinny być interpretowane w kontekście serii eksperymentów. Jeśli $p = 0{,}06$, wówczas badacz powinien zadać sobie pytanie, czy efekt jest potencjalnie interesujący, aby go dalej badać. Fisher[5] zawsze usiłował poprawić plan eksperymentu, gdy wartości p mieściły się między 0,05 i 0,2.

Z tych dyskusji wynika jedno: nic – nawet najlepsze techniki obliczeniowe – i nikt nigdy nie zwalnia badacza z krytycznego myślenia na temat uzyskanych wyników, zarówno w terminach statystycznych, jak i merytorycznych. Powtarzam to zdanie jak mantrę, ale krytyczne myślenie jest kluczem poprawnego wnioskowania.

5 „We feel that p values should be interpreted in the context of a series of experiments. If $p = 0.06$, then the researcher should ask if the effect is of potential interest to explore further. Fisher always attempted to improve the design when p values were between 0.05 and 0.2".

Rozdział 2. „Uzależnienie" od oprogramowania

2.1. Wprowadzenie

Dzięki rozwojowi technik obliczeniowych, zarówno w sferze sprzętu, jak i oprogramowania, obecnie analizy statystyczne wyników badań przeprowadzane są w komputerach, przy użyciu konkretnego pakietu statystycznego. Programy różnią się między sobą – nie są one takie same np. w zakresie zaimplementowanych metod statystycznych czy konkretnych testów.

Chciałbym, abyśmy zastanowili się nad problemem założeń stosowanych metod statystycznych, a także nad ograniczeniami możliwości weryfikacji założeń tych metod, spowodowanych wyborem tego, a nie innego oprogramowania statystycznego. Zderzenie tych dwóch problemów jeszcze dosadniej pokaże rozziew między teorią statystyki i jej stosowaniem.

Rozważania te poprowadzę, wykorzystując metody analizy wariancji, stosunkowo dobrze znane badaczom w naukach społecznych i medycznych. Rozpocznę od jednoczynnikowej jednozmiennowej analizy wariancji. W metodzie tej testujemy zagadnienie:

$$\begin{cases} H_0: \mu_1 = \mu_2 = \cdots = \mu_k \\ H_1: \neg(\mu_1 = \mu_2 = \cdots = \mu_k) \end{cases} \tag{2.1}$$

gdzie μ_i oznacza wartość oczekiwaną badanej zmiennej w i-tej grupie; $i = 1, 2, \ldots, k$.

W hipotezie zerowej sugerujemy, że wszystkie wartości oczekiwane badanej zmiennej w grupach wyznaczonych przez poziomy czynnika są jednakowe. Hipoteza alternatywna jest, jak zwykle, negacją zerowej.

„Powszechnie wiadomo", że u podstaw analizy wariancji leżą następujące założenia:

- badana cecha ma rozkład normalny w każdej z podpopulacji wyznaczonych przez wartości (poziomy) czynnika,
- wariancje badanej cechy w podpopulacjach wyznaczonych przez wartości (poziomy) czynnika są równe (jednorodne).

„Mniej powszechnie wiadomo", że jeszcze powinna być prawdziwa hipoteza zerowa. Przypomnę w tym miejscu to, co napisałem w rozdziale pierwszym. 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 powyższego twierdzenia nie było spełnione. Ale będzie to skutkowało nieprawdziwością tezy. Podobne wnioski będą wynikały z niespełniania każdego z dwóch pozostałych założeń.

Jednak nie będziemy rozważali skutków niespełniania założenia trzeciego – zbyt przypomina to kwadraturę koła, natomiast zajmiemy się dwoma pierwszymi założeniami.

Dowcipnie skomentowali założenia analizy wariancji Khan i Rayner (2003): „Te założenia mogą być naruszone na znacznie więcej sposobów niż mogą być spełnione"[1]. Jednakże założenia mogą być spełnione albo nie, a czy mogą być spełnione na więcej niż jeden sposób? Dopuszczalne są różne odstępstwa od założeń wynikające z odporności (*robustness*) stosowanych metod, ale nie oznacza to ich spełnienia.

2.2. „Założenie normalności"

Wziąłem to sformułowanie w cudzysłów, aby zwrócić uwagę Czytelnika na pewną nieścisłość, niedokładność w rozumieniu tego pojęcia. Stosunkowo często można spotkać zdanie, że założenie normalności dotyczy próbkowego rozkładu prawdopodobieństwa, czyli inaczej mówiąc, rozkładu prawdopodobieństwa cechy w próbie. Na przykład Lantz (2013) pisze:

> Artykuł ten koncentruje się na różnicy między próbkową nienormalnością i populacyjną nienormalnością pod względem oceny istotności statystycznej. Z próbkową nienormalnością mamy do czynienia, gdy dane w próbie mają kształt sugerujący, że ich macierzysta populacja może być nienormalna. Populacyjna nienormalność, z drugiej strony, pojawia się, gdy macierzysta populacja w rzeczywistości jest nienormalna. Z powodu zmienności (wahania, rozproszenia) w próbkach pewne zbiory danych będą charakteryzowane przez próbkę nienormalną, która wskazuje populację nienormalną, pomimo tego, że populacje macierzyste są normalne. Z drugiej strony, z tego samego powodu, inne próbki będą wydawały się względnie normalne, nawet gdy populacje macierzyste są charakteryzowane przez wyraźną nienormalność. W obu przypadkach badacz ma problemy z tym, że nie znając rozkładów macierzystych populacji, może użyć nieoptymalnych metod dla porównania parametrów położenia w tych populacjach[2].

1 „These assumptions can be violated in many more ways than they can be satisfied!".

2 „This paper focuses on the difference between sample non-normality and population non-normality with respect to statistical significance testing. Sample non-normality prevails when the sample data have a shape suggesting that their parent population might be non-normal. Popula-

Zatem, o którym rozkładzie prawdopodobieństwa jest mowa w założeniach analizy wariancji?

Mówienie o normalności rozkładu empirycznego jest całkowicie niepotrzebne, gdyż założenie to jest sprawdzane poprzez wykonanie odpowiedniego testu statystycznego, a jak pamiętamy, wynik testu zawsze dotyczy pewnego aspektu cechy w populacji generalnej.

Jeszcze gorzej, choć już może niewiele, brzmi sformułowanie tego założenia w materiałach do zajęć.

> Założenia dla przeprowadzenia ANOVA:
> 1) populacje są normalne,
> 2) populacje mają takie same (nieznane) wariancje.
> Powyższe warunki są odporne w tym sensie, że można przeprowadzić ANOVA, jeśli populacje są w przybliżeniu normalne (w przeciwnym wypadku stosujemy test Kruskala-Wallisa – test nieparametryczny) i wariancje populacyjne są w przybliżeniu równe (Mehlman, 2017)[3].

Oczywiście „obrońcy swobody wypowiedzi" powołają się na „skrót myślowy". Ale tego typu „skróty myślowe" całkowicie wypaczają sens wypowiedzi i całkowicie dezorganizują próby zrozumienia statystyki u uczących się. Bo cóż w statystyce oznacza „normalność populacji"? Co w ogólności może oznaczać, że populacja jest normalna?

Można także spierać się, czy założenie normalności jest istotnie założeniem w odpowiednim twierdzeniu. W podręczniku Lehmanna (1968) jest ono sformułowane jako twierdzenie dla cechy o rozkładzie normalnym i normalność rozkładu nie jest tam traktowana jako założenie. Jednak takie zróżnicowanie „pojęcia normalności" nie ma wpływu na stosowalność twierdzenia. W praktyce założenie to jest bardzo rzadko spełnione. I co wtedy mamy począć z analizą statystyczną wyników badania? Od razu przechodzimy do nieparametrycznej wersji analizy wariancji, czyli do testu Kruskala-Wallisa? Z jednej strony testy nieparametryczne są słabsze od wersji parametrycznych, tzn. rzadziej uznamy, że istnieją różnice między rozkładami, gdy one rzeczywiście istnieją, niż w przypadku stosowania

---

tion non-normality, on the other hand, prevails when a parent population actually is non-normal. Owing to the variations in samples, some sets of data will be characterized by sample non-normality which indicates population non-normality even though the parent populations are normal. On the other hand, for the same reason, other samples will seem relatively normal even though the parent populations are characterized by a distinct non-normality. The problem is that in both cases a researcher with no previous understanding of the distribution of the parent populations may use suboptimal methods to compare locations from these populations".

3 „Assumptions for doing ANOVA:
1) the populations are normal,
2) the populations have same (unknown) variance.
The above conditions are robust in the sense one can use ANOVA if the populations are approximately normal (otherwise the Kruskal-Wallis Test – a nonparametric test) and the population variances are approximately equal".

metod parametrycznych. Z drugiej zaś strony, dla jednoczynnikowej analizy wariancji istnieje test Kruskala-Wallisa, lecz cóż zrobić w sytuacji np. dwuczynnikowej analizy wariancji czy jednoczynnikowej analizy kowariancji, dla których brakuje nieparametrycznych odpowiedników?

Statystycy od lat 40. XX w. próbują uwolnić się z gorsetu normalności rozkładu, z różnym skutkiem, ale mimo wszystko bardzo często możemy już zrezygnować z ortodoksji tego założenia. Jednak opowiadanie – krótkie – o zmaganiach z normalnością rozkładu, chciałbym zacząć od względnie nowego artykułu Wilcoxa (2002).

> W wielu książkach omawiających stosowanie statystyki ciągle twierdzi się, że gdy pracuje się na średnich, „nienormalność" rozkładu nie budzi poważnego zaniepokojenia, z wyjątkiem być może sytuacji, gdy rozmiar próbki jest bardzo mały. Te stwierdzenia nie są oparte na niedorzecznych spekulacjach, ale teraz już wiemy, że taki punkt widzenia jest niepoprawny i rozumiemy już, dlaczego we wcześniejszych badaniach nie zauważano poważnych problemów związanych z konwencjonalnymi technikami. W pewnych realistycznych sytuacjach „nienormalność" budzi niepokój nawet dla próbek o liczebności 300 i pewne problemy nie tracą znaczenia bez względu na to, z jak liczną próbką mamy do czynienia[4].

Takie opowiadanie, jacy to jesteśmy już mądrzy i jak głupio kiedyś prowadzono analizę jest całkowicie bezsensowne tak długo, jak długo nie będziemy dysponować skutecznymi i prostymi (przynajmniej względnie prostymi) narzędziami do realizacji tych wzniosłych idei.

Nim przejdziemy do analizowania przykładów, poświęćmy chwilę na zastanowienie się nad drugim założeniem analiz wariancji, tj. założeniem o jednorodności (równości) wariancji badanej cechy w podpopulacjach. To założenie jest sprawdzane za pomocą odpowiedniego testu w czasie realizacji obliczeń. Na przykład w SPSS-ie jest to test Levene'a. A jakie założenia leżą u podstaw testu Levene'a? Czy założenie normalności rozkładu badanej cechy ma jakiś wpływ na stosowalność testu umożliwiającego ocenę jednorodności wariancji? Czy test Levene'a to jedyny test do oceny jednorodności wariancji?

Istnieje kilka testów do oceny jednorodności wariancji, np. oryginalny test Levene'a, w którym wykorzystywane są średnie (Levene, 1960), zmodyfikowany test Levene'a zbudowany na medianach (Brown, Forsythe, 1974), testy Levene'a z korekcją Welcha (Welch, 1951) i kilka innych. Dlaczego w SPSS-ie zaimplementowany jest oryginalny test Levene'a? Jak zauważa Constance A. Mara w swojej dysertacji doktorskiej (Mara, 2013), mimo że test ten nie jest polecany w literaturze (np. Conover i wsp., 1981; Lim, Loh, 1996), to jest ciągle implementowany w popularnych programach statystycznych.

4 „Many applied statistics books still claim that when working with means, nonnormality is not a serious concern except possibly when sample sizes are very small. These claims are not based on wild speculations, but it is now known that this view is incorrect, and we understand why earlier studies missed serious problems with conventional techniques. In some realistic situations, nonnormality is a concern even with a sample size of 300, and some problems persist no matter how large the sample size might be!".

PRZYKŁAD 2.1

Przyjrzyjmy się pewnym szczegółom. Przedstawię wyniki jednoczynnikowej analizy wariancji uzyskane w trzech pakietach statystycznych: SPSS 24, STATA 13 i SYSTAT 13. Oczywiście nie są to wszystkie programy statystyczne istniejące na rynku, ale dla tych trzech mam licencje. Nie chodzi mi o porównywanie tych programów, tylko o zwrócenie uwagi, że użycie konkretnego programu statystycznego determinuje wybór – z konieczności – możliwych do wykorzystania testów statystycznych, gdyż tylko takie są tam zaimplementowane.

Dane będące podstawą tego przykładu pochodzą z porównania efektywności trzech terapii leczenia „łokcia tenisisty", ocenianej po ośmiu tygodniach od zakończenia terapii (Kubot, 2017).

$$\text{Kod\_grupy3} = \begin{cases} 0 & \text{grupa kontrolna (terapia „klasyczna")} \\ 1 & \text{terapia falą uderzeniową} \\ 2 & \text{terapia ultradźwiękami} \end{cases}$$

DASH_8 – ocena funkcji bolesnej kończyny (im wyższa wartość wskaźnika, tym gorsze funkcjonowanie) po ośmiu tygodniach po terapii. DASH_8 jest zmienną ciągłą.

Pełne wydruki wyników analizy z każdego z trzech programów statystycznych znajdują się w załączniku 2. Tutaj będę posługiwał się wybranymi fragmentami.

Na początku zajmijmy się założeniem normalności rozkładu prawdopodobieństwa badanej cechy, tj. zmiennej DASH_8, w porównywanych trzech grupach pacjentów: poddanych terapii klasycznej, poddanych terapii falą uderzeniową i poddanych terapii ultradźwiękami. W tym momencie abstrahujemy od liczebności próby, która wynosi 120 osób, i formalnie sprawdzamy normalność rozkładu zmiennej DASH_8.

W wyniku użycia programu **SPSS 24** mamy rezultaty przedstawione w tab. 2.1.

Tabela 2.1. Wyniki oceny normalności zmiennej DASH_8 w grupach poddanych odpowiednim terapiom; zastosowane zostały testy Kołmogorowa-Smirnowa i Shapiro-Wilka

Testy normalności rozkładu							
Zmienna	kod_grupy3	Kołmogorow-Smirnow[a]			Shapiro-Wilk		
		Statystyka	*df*	Istotność	Statystyka	*df*	Istotność
DASH_8	0	,108	60	,081	,958	60	**,039**
	1	,156	30	,060	,908	30	**,013**
	2	,105	30	,200*	,962	30	,339

* Dolna granica rzeczywistej istotności.

[a] Z poprawką istotności Lillieforsa.

Pierwsza rzecz, która rzuca się w oczy, to rozbieżność decyzji, które byłyby podjęte na podstawie testu Kołmogorowa-Smirnowa i testu Shapiro-Wilka. Decyzje podjęte na podstawie testu Kołmogorowa-Smirnowa: nie ma podstaw do odrzucenia hipotezy zerowej o normalności rozkładu zmiennej DASH_8 w każdej grupie. Natomiast na podstawie testu Shapiro-Wilka podejmujemy decyzję, że rozkład zmiennej DASH_8 w grupach poddanych terapii klasycznej i falą uderzeniową nie jest rozkładem normalnym. Który test powinniśmy wybrać? Bo oczywiście wygodniejsze dla nas są wyniki testu Kołmogorowa-Smirnowa, ale czy mamy prawo do takiego wyboru?

A oto wyniki uzyskane w programie **STATA 13**.

.by kod_grupy3, sort: swilk DASH_8

Tabela 2.2. Rezultaty oceny normalności zmiennej DASH_8 w poszczególnych grupach poddanych różnym terapiom; zastosowany został test Shapiro-Wilka

-> kod_grupy3 = 0 – terapia klasyczna

Shapiro-Wilk W test for normal data

Variable	Obs	W	V	z	Prob>z
DASH_8	60	0.95874	2.243	1.741	**0.04086**

-> kod_grupy3 = 1 – terapia falą uderzeniową

Shapiro-Wilk W test for normal data

Variable	Obs	W	V	z	Prob>z
DASH_8	30	0.90497	3.020	2.286	**0.01114**

-> kod_grupy3 = 2 – terapia ultradźwiękami

Shapiro-Wilk W test for normal data

Variable	Obs	W	V	z	Prob>z
DASH_8	30	0.96225	1.200	0.377	0.35321

Rozkłady prawdopodobieństwa zmiennej DASH_8 w grupach poddanych terapii klasycznej i falą uderzeniową okazały się nie być normalnymi.

.by kod_grupy3, sort: sfrancia DASH_8, boxcox

Tabela 2.3. Rezultaty oceny normalności zmiennej DASH_8 w poszczególnych grupach poddanych różnym terapiom; zastosowany został test Shapiro-Francia z transformacją Boxa-Coxa

-> kod_grupy3 = 0 – terapia klasyczna

Shapiro-Francia W' test for normal data

Variable	Obs	W'	V'	z	Prob>z
DASH_8	60	0.96837	1.894	1.236	0.10815

```
-> kod_grupy3 = 1 – terapia falą uderzeniową
      Shapiro-Francia W' test for normal data
 Variable | Obs    W'     V'     z    Prob>z
-------------+--------------------------------------------------
  DASH_8 |  30  0.94177   2.044   1.311  0.09488

-> kod_grupy3 = 2 – terapia ultradźwiękami
      Shapiro-Francia W' test for normal data
 Variable | Obs    W'     V'     z    Prob>z
-------------+--------------------------------------------------
  DASH_8 |  30  0.97273   0.957  –0.081  0.53242
```

Zastosowanie testu Shapiro-Francia z transformają Boxa-Coxa doprowadziło do decyzji identycznej z podjętą po teście Kołmogorowa-Smirnowa w SPSS: nie możemy powiedzieć, by którykolwiek z rozkładów nie był rozkładem normalnym.

.by kod_grupy3, sort: sfrancia DASH_8

Tabela 2.4. Rezultaty oceny normalności zmiennej DASH_8 w poszczególnych grupach poddanych różnym terapiom; zastosowany został test Shapiro-Francia, ale bez transformacji Boxa-Coxa

```
-> kod_grupy3 = 0 – terapia klasyczna
      Shapiro-Francia W' test for normal data
 Variable | Obs    W'     V'     z    Prob>z
-------------+--------------------------------------------------
  DASH_8 |  60  0.96837   1.903   1.231  0.10923

-> kod_grupy3 = 1 – terapia falą uderzeniową
      Shapiro-Francia W' test for normal data
 Variable | Obs    W'     V'     z    Prob>z
-------------+--------------------------------------------------
  DASH_8 |  30  0.94177   2.054   1.318  0.09372

-> kod_grupy3 = 2 – terapia ultradźwiękami
      Shapiro-Francia W' test for normal data
 Variable | Obs    W'     V'     z    Prob>z
-------------+--------------------------------------------------
  DASH_8 |  30  0.97273   0.962  –0.072  0.52854
```

W programie STATA mamy do dyspozycji trzy testy normalności rozkładu badanej cechy: test Shapiro-Wilka, test Shapiro-Francia z transformacją Boxa-Coxa (domyślnie jest to transformacja logarytmiczna[5]) i test Shapiro-Francia bez dodatkowych transformacji. Decyzje podejmowane na podstawie testu Shapiro-Wilka oraz testów Shapiro-Francia są różne. Który test wybrać?

5 Zastosowanie transformacji logarytmicznej polega na zlogarytmowaniu wszystkich wartości badanej cechy (tutaj DASH_8) i przeprowadzeniu analizy na danych zlogarytmowanych. Wszystkie relacje między wartościami zarówno obserwowanymi, jak i obliczonymi będą zachowane, gdyż transformacja logarytmiczna jest transformacją różnowartościową.

Test Shapiro-Wilka zaimplementowany w programie STATA jest oparty na oryginalnej propozycji Shapiro i Wilka (1965) (opracowany został dla prób o liczebności do 50 elementów), jednak został poddany pewnej modyfikacji, aby być odpowiednim dla dużych prób (może być używany dla prób o liczebności $4 \leq n \leq 2000$) (Royston, 1982). Test Shapiro-Francia został skonstruowany dla n nieprzekraczających 100 (Shapiro, Francia, 1972), lecz został zmodyfikowany, by być odpowiednim dla prób $5 \leq n \leq 5000$ (Royston, 1983). Oba testy zaimplementowane w STATA nadają się dla dużych prób. Czy próba 120-elementowa to próba duża?

Poniżej rezultaty uzyskanie w programie **SYSTAT 13**.

Tabela 2.5. Wyniki testowania normalności rozkładu zmiennej DASH_8 w próbie (wszystkie terapie zostały połączone)

Test for Normality		
	Test Statistic	p-Value
K-S Test (Lilliefors)	0,063	0,267
Shapiro-Wilk Test	0,985	0,205
Anderson-Darling Test	0,402	> 0,15*

* The p-Value cannot be precisely computed.

Oczywiście nie jest to ocena normalności rozkładu prawdopodobieństwa badanej cechy w porównywanych grupach, lecz ocena normalności w całej próbie. W opisie programu w założeniach analizy wariancji pojawia się tutaj wymaganie symetrii rozkładu badanej cechy w każdej z porównywanych grup zamiast ich normalności.

Podsumowując: co, na podstawie przedstawionych wydruków, możemy powiedzieć na temat spełniania założenia o normalności rozkładu badanej cechy, tutaj DASH_8, w porównywanych grupach terapeutycznych? Mamy rozwiązania alternatywne? Stosunkowo bezpiecznym rozwiązaniem będzie zastosowanie transformacji symetryzującej rozkład prawdopodobieństwa, gdyż po transformacji możemy użyć odpowiednich testów i podjąć decyzję odnoszącą się do populacji generalnej. Można także, przy odpowiednio licznej próbie, skorzystać z twierdzeń granicznych. „Najsłabszym" rozwiązaniem będzie ocena symetrii rozkładu na podstawie współczynnika skośności obliczonego z próby. Będziemy tu mieli do czynienia z oceną rozkładu w próbie, a nie w populacji generalnej.

2.3. Założenie jednorodności wariancji

Przyjrzyjmy się teraz sposobom sprawdzania w wybranych programach statystycznych założenia o jednorodności wariancji badanej cechy w porównywanych grupach.

SPSS 24

Tabela 2.6. Wynik testu Levene'a dla zmiennej DASH_8

Test jednorodności wariancji[6]			
DASH_8			
Test Levene'a	*df*1	*df*2	Istotność
10,737	2	117	,000

STATA 13

Tabela 2.7. Rezultat testu Bartletta równości wariancji zmiennej DASH_8

Bartlett's test for equal variances: chi2(2) = 15.2007 Prob>chi2 = 0.001

SYSTAT 13

Tabela 2.8. Rezultaty testów Levene'a dla zmiennej DASH_8

Levene's Test for Homogeneity of Variances		
	Test Statistic	p-Value
Based on Mean	10,737	0,000
Based on Median	10,049	0,000

Mimo różnych testów wykorzystanych do sprawdzania jednorodności wariancji decyzja w każdym przypadku jest taka sama: wariancje zmiennej DASH_8 w porównywanych grupach nie są jednorodne (nie są równe).

Jednak warto w tym miejscu zadać sobie pytanie: jaki jest wpływ ewentualnego niespełniania założenia o normalności rozkładów prawdopodobieństwa badanej cechy w porównywanych grupach na wyniki testów jednorodności wariancji?

Lim i Loh (1996) porównują własności siedmiu testów oceny jednorodności wariancji i ich bootstrapowych[7] wersji dla małych oraz umiarkowanie dużych

6 W najnowszej wersji SPSS 25 (inna nazwa to PS IMAGO 5.0) zaimplementowane są już cztery wersje testu Levene'a: bazujący na średniej, bazujący na medianie, bazujący na medianie i skorygowanej liczbie stopni swobody oraz bazujący na średniej obciętej.

7 Metody bootstrapowe polegają na wylosowaniu próby, najlepiej reprezentatywnej, z populacji generalnej, a następnie losowanie **ze zwracaniem** próbek o tej samej liczebności, co wylosowana wcześniej próba. Próbki ze zwracaniem możemy losować dowolną liczbę razy. Oczywiście w próbkach losowanych ze zwracaniem możemy mieć po kilka tych samych osób i oczywiście te próbki nie będą już próbkami reprezentatywnymi. Interesuje nas pewna statystyka, której rozkład i wartość możemy oszacować na podstawie próbki reprezentatywnej. W ten sam sposób szacujemy rozkład albo wartości statystyki na podstawie próbek bootstrapowych i wykorzystujemy je do rozwiązania zagadnienia

wielkości próby. Oceniają ich odporność pod względem rzeczywistego poziomu istotności i mocy w symulowanym eksperymencie. Wśród tych siedmiu testów znalazły się testy Levene'a, test Bartletta, test Boxa-Andersena i jackknife test. Ich ocena: test Levene'a i jeden z testów jackknife są najodporniejsze pod względem poziomu istotności, zaś bootstrapowa wersja testu Levene'a ma największą moc. Lim i Loh sugerują jednak, aby do tych wyników podchodzić z dużą ostrożnością, gdyż zostały one uzyskane na podstawie jednego symulowanego zbioru danych.

Vorapongsathorn i wsp. (2004), porównując prawdopodobieństwo błędu pierwszego rodzaju i moc trzech testów: Bartletta, Levene'a i Cochrana, stwierdzili, że o ile test Bartletta był wrażliwy na niespełnianie założenia o normalności, to testy Cochrana i Levene'a były odporne, gdy założenie to nie było spełnione. Ponadto test Levene'a był „całkiem dobry" (cokolwiek by to miało znaczyć) dla małych prób i grup równolicznych. Pod względem mocy test Bartletta miał największą moc we wszystkich sytuacjach. Z kolei w przypadku, gdy jedna z wariancji była bardzo duża za najlepszy uznano test Cochrana.

Nordstokke i Zumbo (2007) we wstępie do swojej pracy zwracają uwagę na fakt sporej liczby różnych testów Levene'a służących ocenie jednorodności wariancji; w tej chwili jest to już rodzina testów, a nie jeden test. Elementy tej rodziny mają różne własności. Na podstawie analizy własności testu Levene'a dla dwóch wariancji, wykorzystującego średnie (oryginalny test Levene'a (1960)), autorzy ocenili, że test ten jest wrażliwy na „nienormalność" rozkładu prawdopodobieństwa badanej cechy.

Isabel Parra-Frutos (2009) porównywała odporność na „nienormalność" rozkładu badanej cechy pod kątem poziomu istotności i mocy testu trzy modyfikacje testu Levene'a: modyfikację Keyesa-Levy'ego, test z poprawką Satterhwaite'a oraz test Levene'a z obiema poprawkami jednocześnie. W tym przypadku problem polega na braku tych testów w prezentowanych programach statystycznych. Potwierdza to tylko fakt, iż test Levene'a jest bardzo bogatą rodziną testów oceny jednorodności wariancji.

Często autorzy porównujący testy jednorodności wariancji pod względem odporności na „nienormalność" rozkładu badanej cechy, pod względem poziomu istotności czy mocy testu opowiadają o rzeczach nieosiągalnych dla użytkownika konkretnego pakietu statystycznego. Na przykład Gastwirth i wsp. (2009) porównuje kilka modyfikacji testu Levene'a, zaś Lee i wsp. (2010) stwierdzają w podsumowaniu, że byłoby wysoce pożyteczne, aby badacz miał możliwość skorzystać z wyników kilku testów oceny jednorodności wariancji.

Warto tu jeszcze wspomnieć o wyjątkowo prostym do obliczeń i interpretacji teście F_{max} Hartleya (1950). Lee i wsp. (2010) piszą o nim, że pojawiał się w wielu starszych zaawansowanych podręcznikach statystyki, np. tych autorstwa Winera (1971) czy Kirka (1994). Jednak znalazłem go także w podręczniku Winera i wsp. z 1991 r. Hipoteza zerowa i alternatywna są takie same, jak w przypadku testu Levene'a oraz

(test, estymacja). Nieco więcej informacji na temat metod bootstrapowych, podanych w sposób intuicyjny, można znaleźć na stronach: http://www.jamer.pl/resources/matematyka/bootstrap.pdf oraz https://brain.fuw.edu.pl/edu/index.php/WnioskowanieStatystyczne/Bootstrap (dostęp: 30.05.2011).

innych testów jednorodności wariancji, a statystyka będąca podstawą testu to po prostu iloraz największej wariancji i najmniejszej wariancji w podgrupach.

$$F_{max} = \frac{s^2_{max}}{s^2_{min}} \tag{2.2}$$

Tablice wartości krytycznych opracowane zostały przez Hartleya tylko dla równolicznych grup. Brak uogólnienia tego testu na grupy o nierównych liczebnościach.

Mehlman (2017) w swych materiałach do zajęć wspomina o często wykorzystywanej konwencji:

> Konwencja określająca regułę pozwalającą uznać wariancje za równe. Jeśli największe próbkowe odchylenie standardowe jest większe nie więcej niż dwa razy od najmniejszego odchylenia standardowego, można używać technik ANOVA przy założeniu, że wariancje są takie same. W niektórych podręcznikach warunek ten jest wyrażony jako czterokrotność dla wariancji[8].

Można też spotkać warunek, że jeśli krotność między największym odchyleniem standardowym i najmniejszym nie przekracza 3 (albo 10 w przypadku wariancji), to możemy korzystać z metod analizy wariancji. Co prawda, odnosi się to do próby, a nie populacji generalnej.

Czy w takiej sytuacji w ogóle potrafimy rozstrzygnąć, czy założenie o jednorodności wariancji jest spełnione czy też nie? A może powinniśmy zwrócić uwagę na liczebność podgrup? Jak zauważył Zimmerman (2004):

> Mimo że w wielu współczesnych podręcznikach stwierdza się, że wstępne testy nie są już preferowane i tak naprawdę niekonieczne, wielu badaczy w psychologii, edukacji i innych naukach społecznych nie zdaje sobie sprawy z poważnych wad tych testów. Aktualne rekomendacje nie dostarczają informacji, że one znacząco modyfikują poziom istotności. Wszystkie konkluzje wynikające z obecnej pracy sugerują, że badacze powinni zwracać większą uwagę na różnice w liczebnościach próby jako na sygnał niebezpieczeństwa bardziej niż na niejednorodność wariancji[9].

Nie było moim celem przedstawienie łatwego przepisu na stosowanie jednoczynnikowej analizy wariancji, lecz zwrócenie uwagi Czytelnika na bardzo skomplikowane powiązania między założeniami i testami służącymi do ich „sprawdzania".

8 „Convention: rule for establishing equal variance. If the largest sample standard deviation is less than twice the smallest sample standard deviation, one can use ANOVA techniques under the assumption the variances are all the same. Some textbooks use four times the smallest sample variance instead of just twice".

9 „Although many current textbooks state that preliminary tests are no longer favoured and not really necessary, many researchers in psychology, education and other social sciences are unaware of the serious disadvantages of these tests. Current recommendations do not convey the message that they substantially modify the significance level. All findings in the present note suggest that researchers should pay more attention to differences in sample sizes as a danger signal rather than to heterogeneity of variance".

Ale pozostawienie Czytelnika z mnóstwem pytań, na które najprawdopodobniej nie potrafi znaleźć satysfakcjonującej odpowiedzi (co prawda, w Internecie znajdzie się prosta rada na wszystko) też raczej nie jest dobrym rozwiązaniem. Zatem spróbuję, tak prosto jak to będzie możliwe i jak będę potrafił, wyjaśnić – również sobie – niektóre z zasygnalizowanych problemów.

Wróćmy do założenia normalności rozkładu badanej cechy w porównywanych grupach. Co dzieje się z wynikami analizy wariancji, gdy założenie to nie jest spełnione? Jak dalece to założenie może być niespełnione? Choć ostatnie pytanie może wydawać się absurdalne (założenie jest albo nie jest spełnione), to w praktyce wcale nie jest takie bezsensowne. Powodem tego jest – mniejsza lub większa – odporność statystyk będących podstawą różnych testów na niespełnianie założeń formułowanych w odpowiednich twierdzeniach statystyki teoretycznej. Jak dalece statystyka *F*-Snedecora, stanowiąca podstawę większości testów w analizach wariancji, jest odporna na niespełnianie założenia normalności rozkładu? Należałoby jednak w tym miejscu sprecyzować, jakiego aspektu statystyki czy testu dotyczy owa odporność. Może ona dotyczyć rzeczywistego poziomu istotności testu, mocy testu, a także wartości statystyki.

Jak zwykle w przypadku problemów ze stosowaniem metod statystycznych nie ma na to prostej i jednoznacznej odpowiedzi. Co więcej, samo sformułowanie założeń i problemów może prowadzić do konfuzji. Na przykład: „W większości podręczników statystyki jednoczynnikowa analiza wariancji o efektach stałych (ANOVA) rekomendowana jest jako najlepsza metoda porównywania wartości oczekiwanych kilku populacji, jeśli mają one w przybliżeniu rozkłady normalne i mają podobne wariancje" (Lantz, 2013)[10]. Natychmiast rodzi się pytanie: jak mocno ma być przybliżony do normalnego rozkład badanej cechy, w jakich aspektach oceniamy przybliżenie: kształtu wykresu funkcji gęstości, parametrów charakteryzujących rozkład, czy też jeszcze inaczej? Podobnie jest z podobieństwem wariancji.

A oto przykłady odpowiedzi na pytanie o odporność analizy wariancji. Mówiąc precyzyjniej, pytanie będzie dotyczyło odporności statystyki *F*-Snedecora, która jest podstawą testu oceny równości wartości oczekiwanych.

Glass i wsp. (1972) – w oryginale „skośność populacji" (poprawniej byłoby napisać o skośności rozkładu badanej cechy w populacji generalnej) ma bardzo mały wpływ zarówno na poziom istotności, jak i moc testu w modelach o efektach stałych. Spłaszczenie (kurtoza) rozkładu ma drobny wpływ na nominalny poziom istotności zarówno w przypadku grup równolicznych, jak i nierównolicznych. Efekt spłaszczenia rozkładu może być znaczący dla małych *n* (liczebności w grupach).

Bock (1975) – nawet jeśli rozkłady badanych cech znacznie odbiegają od rozkładu normalnego, suma 50 albo więcej obserwacji ma w przybliżeniu (znów w przybliżeniu!) rozkład normalny. Dla rozkładów mniej „nienormalnych" przybliżenie jest już dobre dla 10 albo 20 obserwacji.

10 „Most statistics textbooks recommend the one-way fixed effect analysis of variance (ANOVA) as the best method for comparing the means of several populations if they are approximately normally distributed and have similar variances".

Babu i wsp. (1999) – autorzy proponują rozwiązanie stosowalne (właściwe) zarówno w sytuacjach jednorodności (*homoscedastic*), jak i niejednorodności wariancji (*heteroscedastic*), zakładając, że rozkłady badanych cech są symetryczne. Jednak rozwiązania te są oparte albo na 15% przyciętych (*trimmed*) średnich, albo na próbkowych medianach, których wartości są estymowane metodą bootstrap.

Awan (2001) – efekt „nienormalności" zmniejsza się wraz ze wzrostem liczebności w grupie, natomiast ten efekt zwiększa się wraz ze wzrostem liczby grup.

Garcia-Perez (2008) – autor proponuje pewne przybliżenia statystyki *F*, które są użyteczne w badaniach odporności pod względem poziomu istotności i wartości krytycznej, gdy rozkład w modelu nieznacznie odbiega od rozkładu normalnego (np. jest zniekształconym normalnym).

Khan i Rayner (2003) – autorzy od razu proponują przejście do testu Kruskala-Wallisa, czyli nieparametrycznej wersji jednoczynnikowej analizy wariancji, gdy nie jest spełnione założenie o normalności rozkładu badanej cechy.

2.4. Testy porównań wielokrotnych

Kolejne istotne pytanie związane z założeniem normalności badanej cechy: jak niespełnianie tego założenia wpływa na możliwość stosowania testów porównań wielokrotnych? Poniżej przedstawiam fragmenty wydruków z trzech porównywanych programów.

SPSS 24

Tabela 2.9. Wyniki testów *post hoc* (testów porównań wielokrotnych) dla zmiennej DASH_8

Porównania wielokrotne						95% przedział ufności	
Testy	(I) kod_ grupy3	(J) kod_ grupy3	różnica średnich (I–J)	błąd standardowy	istotność	dolna granica	górna granica
Test Bonferroniego	,00	1,00	27,65508*	4,79321	,000	16,0129	39,2973
		2,00	9,72760	4,79321	,134	–1,9146	21,3698
	1,00	,00	–27,65508*	4,79321	,000	–39,2973	–16,0129
		2,00	–17,92749*	5,53472	,005	–31,3707	–4,4843
	2,00	,00	–9,72760	4,79321	,134	–21,3698	1,9146
		1,00	17,92749*	5,53472	,005	4,4843	31,3707

Tabela 2.9 (cd.)

Porównania wielokrotne						95% przedział ufności	
Testy	(I) kod_ grupy3	(J) kod_ grupy3	różnica średnich (I–J)	błąd standardowy	istotność	dolna granica	górna granica
Test Dunnetta T3	,00	1,00	27,65508*	4,02787	,000	17,8579	37,4523
		2,00	9,72760	4,83458	,136	–2,0761	21,5313
	1,00	,00	–27,65508*	4,02787	,000	–37,4523	–17,8579
		2,00	–17,92749*	4,25902	,000	–28,4336	–7,4214
	2,00	,00	–9,72760	4,83458	,136	–21,5313	2,0761
		1,00	17,92749*	4,25902	,000	7,4214	28,4336

* Różnica średnich jest istotna na poziomie 0.05

W przypadku programu SPSS mamy przedstawione wyniki dwóch testów: testu porównań wielokrotnych, wymagającego spełnienia założenia o jednorodności wariancji (test Bonferroniego) i testu niewymagającego spełnienia tego założenia (test Dunnetta T3).

STATA 13

Tabela 2.10. Rezultaty testów porównań wielokrotnych uzyskane w programie STATA 13

```
          (Bonferroni)
Row Mean-|
Col Mean |    0     1
---------+----------------------
   1 |  -27.6551
     |   0.000
     |
   2 |  -9.7276  17.9275
     |   0.134   0.005
```

```
      Comparison of DASH_8 by kod_grupy3
          (Scheffe)
Row Mean-|
Col Mean |    0     1
---------+----------------------
   1 |  -27.6551
     |   0.000
     |
   2 |  -9.7276  17.9275
     |   0.132   0.007
```

```
        Comparison of DASH_8 by kod_grupy3
              (Sidak)
Row Mean-|
Col Mean |     0     1
---------+----------------------
   1 |  -27.6551
     |    0.000
     |
   2 |  -9.7276  17.9275
     |    0.128   0.005
```

W programie STATA 13 mamy jedynie możliwość wyboru jednego lub więcej testów porównań wielokrotnych spośród trzech testów wymagających spełnienia założenia o jednorodności wariancji.

SYSTAT 13

▼ Hypothesis Tests
Post Hoc Test of DASH_8
Using least squares means.
Using separate variances error terms.

Tabela 2.11. Wyniki testu Dunnetta T3 porównań wielokrotnych w programie SYSTAT 13

Dunnett's T3 Test					
KOD_GRUPY3(i)	KOD_GRUPY3(j)	Difference	p-Value	95% Confidence Interval	
				Lower	Upper
0,000	1,000	27,655	0,000	17,858	37,452
0,000	2,000	9,728	0,136	–2,076	21,531
1,000	2,000	–17,927	0,000	–28,434	–7,421

> POST KOD_GRUPY3 / T3.

W programie SYSTAT 13 wybieramy testy porównań wielokrotnych w zależności od wyniku testu Levene'a oceny jednorodności wariancji. Grupy testów są podobne, jak w SPSS 24, jednak nie możemy tutaj wybrać jednocześnie testu wymagającego jednorodności wariancji i testu niewymagającego spełnienia tego założenia.

W przypadku przedstawionych wyników testów porównań wielokrotnych decyzje podjęte na ich podstawie są takie same: wartość oczekiwana zmiennej DASH_8 w grupie poddanej terapii falą uderzeniową jest istotnie mniejsza niż w grupach dwóch pozostałych terapii. Wartości oczekiwane grup: poddanych terapii klasycznej i terapii ultradźwiękami „nie różnią się między sobą" (tab. 2.11).

Dodatkowo warto zauważyć, że nie wszystkie z zastosowanych testów porównań wielokrotnych zostały użyte poprawnie – wszak nie jest spełnione założenie o jednorodności wariancji zmiennej DASH_8 w porównywanych grupach.

2.5. Normalność a testy porównań wielokrotnych

Zamiast komentarzy związanych z relacją normalności rozkładu badanej cechy i testami porównań wielokrotnych przedstawię tylko kilka cytatów.

Już w 1971 r. Games stwierdził, że „obszar porównań wielokrotnych jest jednym z najbardziej zagmatwanych (dezorientujących) obszarów statystyki i jest tym, który uzyskuje zestaw bardzo zróżnicowanych rekomendacji ze strony autorów wielu tekstów statystycznych o stosowaniu tych metod w naukach behawioralnych”[11] (Games, 1971: 531).

Keselman i wsp. (2002) piszą:

> Liczni autorzy sugerują, że dane zebrane przez badaczy nie mają normalnego kształtu. Zgodnie z metodami oceny porównań wielokrotnych par średnich statystyki tradycyjnej częstym rezultatem jest obciążenie wyników błędem I rodzaju i obniżeniem mocy wykrywania efektów. Jednym z rozwiązań jest otrzymanie wartości krytycznej do oceny istotności statystycznej metodami bootstrap. Do przeprowadzenia krokowych testów bootstrapowych może być użyty system SAS. Autorzy wykorzystali to podejście, gdy ani dane nie miały postaci normalnej, ani nie miały jednakowej zmienności w planach zrównoważonych i niezrównoważonych[12].

Czy jednak metody bootstrapowe są antidotum na wszystkie bolączki z założeniami leżącymi u podstaw stosowanych testów? Metody bootstrapowe również mają swoje ograniczenia, problemy z ich stosowaniem nie wynikają tylko z braku odpowiedniego oprogramowania. Pisze o nich Kochanski (2005): „Dowody matematyczne trafności oszacowań bootstrapowych są poprawne jedynie w granicy przy bardzo dużych próbach”.

Pozostajemy przy „praktycznym”, w najlepszym razie intuicyjnym, wykorzystaniu metod statystycznych:

---

11 „The area of multiple comparisons is one of the more confusing areas of statistics, and is one that receives a widely differing set of recommendations from many applied statistics texts in the behavioral sciences”.

12 „Numerous authors suggest that the data gathered by investigators are not normal in shape. Accordingly, methods for assessing pairwise multiple comparisons of means with traditional statistics will frequently result in biased rates of Type I error and depressed power to detect effects. One solution is to obtain a critical value to assess statistical significance through bootstrap methods. The SAS system can be used to conduct step-down bootstrapped tests. The authors investigated this approach when data were neither normal in form nor equal in variability in balanced and unbalanced designs”.

> Przed przeprowadzeniem jakiejkolwiek analizy statystycznej (np. *t*-test, ANOVA czy analiza korelacji) jest ważne sprawdzenie, czy którekolwiek z założeń wymaganych w pojedynczym teście nie jest naruszone. Wspólnym założeniem jest, aby próbka losowa miała rozkład normalny. […] W wielu analizach statystycznych normalność zakłada się często bezrefleksyjnie, bez jakiegokolwiek dowodu empirycznego czy testu. A w rzeczywistości normalność jest decydująca w wielu parametrycznych metodach statystycznych. […] Gdy takie założenie jest naruszone, interpretacja i wnioskowanie stają się nieuzasadnione[13] (Ahad i wsp., 2011).

2.6. Efekty nieodrzucenia hipotezy zerowej

Niezbędne wydaje się tutaj przypomnienie efektów nieodrzucenia hipotezy zerowej, czyli uznania, że założenie o normalności rozkładu badanej cechy albo o jednorodności wariancji tej cechy w porównywanych grupach zostało spełnione. Ostatnie stwierdzenie, iż założenie zostało spełnione jest nieprawdziwe, gdyż po prostu nie mieliśmy podstaw do odrzucenia hipotezy zerowej, ale to nie znaczy, że mieliśmy prawo ją przyjąć. Wyraźnie widać, że uznawanie na podstawie wyników, iż założenia są spełnione, jest w rzeczywistości iluzoryczne. Racjonalniejsze wydaje się stosowanie praktycznych rozwiązań wykorzystujących odporność odpowiednich statystyk (niewielką asymetrię rozkładów, liczebność próby, porównywanie najmniejszego i największego odchylenia standardowego).

2.7. Podsumowanie

I teraz musimy zadać sobie następujące pytanie: czy mamy prawo użyć jednoczynnikowej analizy wariancji do porównywania wartości oczekiwanych DASH_8 w grupach terapeutycznych? Sądzę, że możemy to zrobić, ale używając praktycznych przesłanek, a więc wykorzystując własności odporności statystyk będących podstawą odpowiednich testów. Spróbujmy przyjrzeć się poszczególnym założeniom.

Założenie normalności rozkładu badanej cechy w porównywanych grupach. Test Shapiro-Wilka jest testem mocniejszym niż test Kołmogorowa-Smirnowa (Razali, Wah, 2011), zatem jego rezultaty powinniśmy wziąć pod uwagę, używając SPSS-a. Co więcej, test Shapiro-Wilka jest polecany dla małych i średnich

13 „Prior to using any statistical analyses (e.g. *t*-test, ANOVA, and correlation) it is important to check that any of the 'assumptions' incurred on individual tests are not violated. A common assumption is that the random sample is normally distributed. […] In many statistical analyses, normality is often conveniently assumed without any empirical evidence or test. Indeed, normality is crucial in many parametric statistical methods. […] When this assumption is violated, the interpretation and inference made be invalid".

pod względem liczebności prób, do 2000. Dla liczniejszych prób test Kołmogorowa-Smirnowa jest rekomendowany przez SAS i innych autorów (Garson, 2012). W psychologii oraz innych naukach społecznych, gdzie próby nie są ogromne, stosowanie testu Shapiro-Wilka byłoby dobrym wyborem. Ale decyzja w wyniku zastosowania testu Shapiro-Wilka brzmi: rozkłady zmiennej DASH_8 nie są rozkładami normalnymi. Może więc powinniśmy wybrać inną drogę postępowania? Możemy skorzystać z przesłanek Bocka (1975) dotyczących liczebności próby albo przesłanek Glassa i wsp. (1972), pozwalających wykorzystać symetrię rozkładu. Współczynniki skośności z próby mieszczą się w przedziale [–1; 1] (załącznik 2), zatem taka niewielka asymetria rozkładów nie powinna zniekształcać wyników analizy. Liczebność próby wynosi 120 osób, a więc jest to próba na tyle duża, abyśmy mogli korzystać z twierdzeń granicznych.

Założenie jednorodności wariancji badanej cechy w porównywanych grupach. W rozważanych przykładach wykorzystywaliśmy jednoczynnikową analizę wariancji i w tym przypadku brak jednorodności wariancji, przynajmniej w pakiecie SPSS, nie był przeszkodą w zastosowaniu metody – mieliśmy do dyspozycji testy Welcha i Browna-Forsythe'a, czyli testy dla wariancji niejednorodnych. Co robić w przypadku innych pakietów? W STATA13 i SYSTAT13 uzyskujemy informację o braku jednorodności wariancji i możemy ją wykorzystać, stosując jedną z transformacji stabilizujących wariancje, a następnie statystykę *F*-Snedecora dla danych po transformacji. Poniżej zostały przedstawione efekty trzech transformacji (Winer i wsp., 1991):

$$\text{DASH\_8\_tr1: } X'_{ij} = \sqrt{X_{ij}} \tag{2.3}$$

$$\text{DASH\_8\_tr2: } X'_{ij} = \sqrt{X_{ij}} + \sqrt{X_{ij} + 1} \tag{2.4}$$

$$\text{DASH\_8\_tr3: } X'_{ij} = \sqrt{X_{ij} + \frac{1}{2}} \tag{2.5}$$

Uzyskano następujące rezultaty testów jednorodności wariancji zmiennej DASH_8 po transformacji:

SPSS 24

Tabela 2.12. Wyniki testowania jednorodności wariancji, testem Levene'a, zmiennej DASH_8 przed transformacją i po transformacjach zdefiniowanych wzorami (2.3)–(2.5)

Zmienne	Test Levene'a	*df*1	*df*2	Istotność
DASH_8	10,737	2	117	,000
DASH_8_tr1	,733	2	117	,483
DASH_8_tr2	1,062	2	117	,349
DASH_8_tr3	1,231	2	117	,296

STATA 13

Tabela 2.13. Wyniki testowania jednorodności wariancji, testem Bartletta, zmiennej DASH_8 po transformacjach zdefiniowanych wzorami (2.3)–(2.5)

DASH_8_tr1: Bartlett's test for equal variances: chi2(2) = 1.0661 Prob>chi2 = 0.587
DASH_8_tr2: Bartlett's test for equal variances: chi2(2) = 1.6433 Prob>chi2 = 0.440
DASH_8_tr3: Bartlett's test for equal variances: chi2(2) = 1.9607 Prob>chi2 = 0.375

SYSTAT 13

Tabela 2.14. Wyniki oceny jednorodności wariancji zmiennej DASH_8 po transformacji określonej wzorem (2.3)

Levene's Test for Homogeneity of Variances		
	Test Statistic	p-Value
Based on Mean	0,733	0,483
Based on Median	0,512	0,601

Tabela 2.15. Wyniki oceny jednorodności wariancji zmiennej DASH_8 po transformacji określonej wzorem (2.4)

Levene's Test for Homogeneity of Variances		
	Test Statistic	p-Value
Based on Mean	1,062	0,349
Based on Median	0,720	0,489

Tabela 2.16. Wyniki oceny jednorodności wariancji zmiennej DASH_8 po transformacji określonej wzorem (2.5)

Levene's Test for Homogeneity of Variances		
	Test Statistic	p-Value
Based on Mean	1,231	0,296
Based on Median	0,827	0,440

Po każdej z tych transformacji nie mamy podstaw do odrzucenia hipotezy zerowej o jednorodności wariancji. Praktyka prowadzenia analiz statystycznych jest taka, że uznajemy, iż wariancje są jednorodne (czy rzeczywiście założenie jest spełnione?).

Wykorzystując konwencję porównywania największego i najmniejszego odchylenia standardowego w grupach, również uznamy, że wariancje w próbie nie różnią się na tyle, aby uniemożliwić przeprowadzenie jednoczynnikowej analizy wariancji w wersji parametrycznej. A co dzieje się w populacji generalnej?

Tabela 2.17. Statystyki opisowe zmiennej DASH_8 w grupach poddanych różnym terapiom

	N	Średnia	Odchylenie standardowe	Błąd standardowy	95% przedział ufności dla średniej		Minimum	Maksimum
					dolna granica	górna granica		
,00	60	42,1370	25,37181	3,27549	35,5827	48,6912	,00	90,15
1,00	30	14,4819	12,83932	2,34413	9,6876	19,2762	,00	43,97
2,00	30	32,4094	19,47637	3,55588	25,1368	39,6820	,00	69,82
Ogółem	120	32,7913	24,09070	2,19917	28,4367	37,1459	,00	90,15

Największe odchylenie standardowe to 25,37, zaś najmniejsze 12,84; ich iloraz wynosi 25,37/12,84 = 1,98 < 2, a więc jest mniejszy od 3.

Transformacje stabilizujące wariancje jednak mają jedno, ogromnie ważne zastosowanie w sytuacjach, gdy w używanym programie statystycznym nie zaimplementowano testów porównań wielokrotnych, które nie wymagają równości wariancji, jak w STATA. Dla danych po transformacji możemy używać testów Bonferroniego, Scheffego czy Šidaka; wszystkie zastosowane transformacje są transformacjami różnowartościowymi.

Rozdział 3. Moc testu statystycznego

3.1. Wprowadzenie

W tym rozdziale ograniczymy się do rozważenia zagadnień związanych z mocą testu jedynie w teorii Neymana-Pearsona, która jest praktycznie jedyną stosowaną w naukach społecznych teorią testowania hipotez statystycznych w paradygmacie częstościowym. W tej teorii oprócz hipotezy zerowej wymagana jest hipoteza alternatywna. Jakie są relacje między tymi hipotezami? Skąd bierze się hipoteza alternatywna?

Jak pamiętamy, hipoteza statystyczna to każde przypuszczenie dotyczące rozkładu prawdopodobieństwa badanej statystyki, a test statystyczny jest procedurą umożliwiającą badaczowi podjęcie decyzji o prawdziwości bądź fałszywości testowanej hipotezy. Umożliwia podjęcie decyzji, lecz nie pozwala rozstrzygnąć o prawdziwości albo fałszywości hipotezy statystycznej. Ta zasadnicza różnica między podjęciem decyzji a obiektywnym rozstrzygnięciem bardzo częsta umyka uwadze osób stosujących metody statystyczne.

Hipoteza zerowa (H_0) zarówno w teorii Fishera, jak i teorii testowania hipotez Neymana-Pearsona jest hipotezą statystyczną, która odpowiada skonstruowanemu modelowi badawczemu. Pojęcie hipotezy alternatywnej funkcjonuje tylko w teorii Neymana-Pearsona – jest ona dopełnieniem hipotezy zerowej do całego zbioru hipotez możliwych do sformułowania w analizowanym zagadnieniu. Na przykład porównując wartości oczekiwane pewnej cechy w dwóch grupach badanych elementów (oczywiście nie muszą to być ludzie), formułujemy hipotezy:

$$\begin{cases} H_0: \mu_1 = \mu_2 \\ H_1: \mu_1 \neq \mu_2 \end{cases} \tag{3.1}$$

Oczywiste jest, że dwie wartości oczekiwane mogą albo być równe, albo nie, i nie ma tu więcej możliwości. Suma mnogościowa hipotez H_0 i H_1 wyczerpuje wszystkie możliwe relacje między tymi dwoma parametrami. Analogicznie, porównując dwa rozkłady prawdopodobieństwa, formułujemy hipotezy:

$$\begin{cases} H_0: F_1 = F_2 \\ H_1: F_1 \neq F_2 \end{cases} \tag{3.2}$$

I znów suma mnogościowa H_0 i H_1 zawiera wszystkie możliwe relacje między F_1 i F_2. Sytuacja jest analogiczna w przypadku większej liczby wartości oczekiwanych albo rozkładów:

$$\begin{cases} H_0: \mu_1 = \mu_2 = \cdots = \mu_k \\ H_1: \neg(\mu_1 = \mu_2 = \cdots = \mu_k) \end{cases} \text{ albo } \begin{cases} H_0: F_1 = F_2 = \cdots = F_k \\ H_1: \neg(F_1 = F_2 = \cdots = F_k) \end{cases} \tag{3.3}$$

Co prawda, sformułowanie hipotezy alternatywnej jest nieco inne niż w przypadku dwóch wartości oczekiwanych (prawo de Morgana). Ale równie dobrze w sytuacji porównywania dwóch wartości oczekiwanych albo rozkładów możemy napisać: $H_1: \neg(\mu_1 = \mu_2)$; $H_1: \neg(F_1 = F_2)$.

W przypadku hipotez dotyczących współczynnika korelacji liniowej mamy:

$$\begin{cases} H_0: \rho = 0 \\ H_1: \rho \neq 0 \end{cases} \tag{3.4}$$

Jakie są relacje między hipotezami: zerową i alternatywną? Otóż, prawdziwa może być dokładnie jedna z nich i mając do dyspozycji ułomne narzędzia testowania hipotez, usiłujemy rozstrzygnąć, która z nich mogłaby być prawdziwa. Jak dalece nam się to nie udaje, próbuję pokazać w niniejszej książce.

Wróćmy jednak do zagadnienia mocy testu. Ale najpierw przypomnienie pojęcia błędu drugiego rodzaju i jego prawdopodobieństwa. Błąd drugiego rodzaju polega na przyjęciu fałszywej hipotezy zerowej i jego prawdopodobieństwo oznaczane jest przez β, a moc testu to $1 - \beta$. Jednak musimy pamiętać, że wartość β zależy od konkretnej hipotezy alternatywnej (funkcja mocy testu). W tym kontekście przez moc testu będziemy rozumieć zdolność testu (w terminach prawdopodobieństwa) do wykrycia fałszywości hipotezy zerowej w sytuacji, gdy jest ona rzeczywiście fałszywa. Jeśli formułujemy prostą hipotezę zerową i prostą hipotezę alternatywną, to mamy do czynienia tylko z dwupunktowym zbiorem parametrów wyznaczających dwa rozkłady prawdopodobieństwa i hipoteza alternatywna jest jednoznacznie określona. Ale w przypadku złożonej hipotezy alternatywnej zbiór parametrów jest znacznie bogatszy, często ma on moc continuum (moc zbioru liczb rzeczywistych). Zamiast mówić o mocy testu używa się więc 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:

$$\begin{aligned} \alpha &= \pi(H_0) = Pr(\text{odrzucenie } H_0 | H_0 \text{ jest prawdziwa}) \\ 1-\beta &= \pi(H_1) = Pr(\text{odrzucenie } H_0 | H_0 \text{ jest fałszywa}) \end{aligned} \tag{3.5}$$

Natomiast w sytuacji hipotez złożonych:

$$\begin{aligned} \alpha &= \max_{\theta \text{ w} H_0} P_0(\text{odrzucić } H_0) = \max_{\theta \text{ w} H_0} \pi(\theta) \\ \beta &= \max_{\theta \text{ w} H_1} [1-\pi(\theta)] \end{aligned} \tag{3.6}$$

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

Rasch (2012) formułuje to dosadniej:

> Moc testu jest w statystyce matematycznej definiowana jako prawdopodobieństwo $\pi(\theta)$ odrzucenia hipotezy zerowej jako funkcji parametru θ, a nie jak wyrażone to jest na przykład u Leventhala i Huynha (1996) i w wielu innych psychologicznych tekstach, jako prawdopodobieństwo odrzucenia niepoprawnej hipotezy zerowej. Zatem moc jest definiowana jako:

$$\pi(\theta) = \begin{matrix} \alpha(\theta) \text{ dla } \theta \in \Omega_0 \\ 1-\beta(\theta) \text{ dla } \theta \in \Omega_A \end{matrix} \text{ a nie jako: } 1-\beta(\theta) \text{ dla } \theta \in \Omega_A \text{[1]} \tag{3.7}$$

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

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases} \tag{3.8}$$

Najogólniej mówiąc, własności testu, w tym także jego moc, będą zależały od prawdziwej wartości parametru w H_1, a tej nie znamy. I mamy problem. Jego

1 „The power of a test is in mathematical statistics defined as the probability $\pi(\theta)$ of rejecting the null hypothesis as a function of θ, and not as stated for instance by Leventhal and Huynh (1996) and in many other psychological texts, as the probability of rejecting a wrong null hypothesis. Thus power is defined as: $\pi(\theta) = \begin{matrix} \alpha(\theta) \text{ for } \theta \in \Omega_0 \\ 1-\beta(\theta) \text{ for } \theta \in \Omega_A \end{matrix}$ and not as: $1-\beta(\theta)$ for $\theta \in \Omega_A$".

konsekwencją jest konstruowanie testów statystycznych kontrolujących prawdopodobieństwo błędu pierwszego rodzaju, lecz niekontrolujących prawdopodobieństwa błędu drugiego rodzaju. Czytelnika zainteresowanego szczegółami funkcji mocy testu odsyłam do podręcznika Lindgrena (1962). Nie jest to może lektura najprostsza, ale rzetelna.

3.2. 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. Cóż zatem jest obliczane w programach statystycznych? Na to pytanie bardzo trudno sensownie odpowiedzieć. Jak widzieliśmy w rozważaniach teoretycznych, moc testu zależy od wartości parametru zaszytego w hipotezie alternatywnej, której to wartości nie znamy. Williams i Zimmerman (1989) stwierdzają, że

> Moc każdego testu istotności [termin angielski *significance test* nie oznacza tego samego, co polski termin *test istotności* – przyp. W. Sz.] zależy od sposobu wykorzystywania informacji zawartych w próbie, tj. konkretnej statystyki testowej[2], która jest obliczana jako funkcja wartości zaobserwowanych w próbie. Ponadto, moc testu zależy od następujących zmiennych: wielkości próby, wariancji populacyjnej, wielkości różnicy między hipotezą zerową i prawdziwą hipotezą alternatywną, poziomu istotności testu i kierunkowości tego testu[3].

Spośród wymienionych wyżej składowych mocy testu najczęściej znana jest nam jedynie wielkość próby.

Jeżeli chodzi o inne składowe mocy testu, to budzą one sporo zastrzeżeń. Wydaje mi się, że nie wolno nam godzić się na „skróty myślowe" typu „wariancja populacyjna" czy „wielkość różnicy między hipotezą zerową i prawdziwą hipotezą alternatywną". Zgadzając się na nie, powiększamy – i tak wystarczająco duży – zamęt, jaki istnieje w metodach testowania hipotez. Cóż bowiem może oznaczać wariancja populacyjna? Nic. Mówimy jedynie o wariancji badanej cechy w populacji generalnej, a tej nie znamy i najprawdopodobniej nie będziemy znali. Wszak nie przebadamy całej populacji generalnej (bez względu na to, jaka ona jest; wyjątkiem są tutaj tzw. badania wyczerpujące) pod względem analizowanej cechy. A co oznacza sformułowanie o wielkości różnicy między hipotezą zerową i prawdziwą hipotezą alternatywną? Kuriozalny jest zwrot „prawdziwa hipoteza alternatywna"

2 Statystyki będącej podstawą testu (przyp. W. Sz.).

3 „The power of any significance test depends on the way in which the test utilizes information in sample data, that is, the particular test statistic that is calculated as a function of sample values. Furthermore, power depends on the following variables: sample size, population variance, the magnitude of the difference between a null hypothesis and a true alternative hypothesis, significance level, and directionality of the test".

w sytuacji, gdy hipoteza alternatywna jest hipotezą złożoną. Jedynie możemy mówić – ale tylko mówić, gdyż nie potrafimy tego rozstrzygnąć – o prawdziwości jednej z hipotez spośród kontinuum hipotez składających się na hipotezę alternatywną. Co możemy wiedzieć o prawdziwości którejkolwiek z hipotez wchodzących w skład hipotezy alternatywnej? Jak mierzyć odległość między hipotezami, jeśli nie są to prymitywne hipotezy typu:

$$\begin{cases} H_0: \mu = \mu_0 \\ H_1: \mu = \mu_1 \end{cases} \quad ? \tag{3.9}$$

O'Keefe (2007) ujmuje zagadnienie podobnie, choć nieco je 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 stawia pytanie: zakładając, że badacz nie zna wielkości efektu w populacji generalnej, 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. Ale autor zauważa, że „mówienie o *mocy* konkretnego testu może być mylące, ponieważ każdy poszczególny test (tj. o określonej wielkości próby) w rzeczywistości ma wiele różnych wartości mocy"[4].

Hoenig i Heisey (2001) jeszcze raz zauważają od dawna znaną prawidłowość, że 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 i nie umożliwi nam to przyjęcia hipotezy zerowej. Autorzy ci napisali: „Część autorów zauważa, że obserwowana moc może być niespecjalnie przydatna, ale z powodu fatalnej skazy logicznej w naszej wiedzy przechodzi to przeważnie niezauważone"[5].

W piśmiennictwie pojawia się jednak inna ciekawa zależność, której warto poświęcić kilka zdań. Mianowicie, czy istnieje jakiś związek między „brakiem istotności" uzyskanym w teście statystycznym a jakością badania?

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 Seldmeiera i Gigerenzera (1989): *Do Studies of Statistical Power Have an Effect on the Power of Studies*, jest

4 „Thus it can be misleading to speak of *the power* of a given statistical test, because any particular test (i.e. with a specified sample size) actually has many different power values".

5 „A number of authors have noted that observed power may not be especially useful, but to our knowledge a fatal logical flaw has gone largely unnoticed".

nieprecyzyjny, gdyż w artykule jest mowa 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. Oprócz tabel z mocami testów *post hoc* w artykułach tych pojawiają się ciekawe spostrzeżenia, np. jako wynik ograniczeń NHST (*Null Hypothesis Significance Testing*), czyli po prostu testowania hipotez, pewni badacze (np. Carver, 1993) utrzymują, że wielkość efektu, która stanowi miarę praktycznej „ważności"[6], powinna całkowicie zastąpić testowanie hipotez. Jednak przedstawianie i interpretowanie tylko wielkości efektu mogłoby doprowadzić do nadinterpretacji wniosków (Onwuegbuzie, Leech, 2004). Zatem autorzy ci zdają sobie sprawę, że nie ma jednej uniwersalnej metody oceny wielkości (czy faktu istnienia) albo braku istnienia badanych zależności.

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 co do tego, 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. Ale w praktyce to się raczej nie zdarzy, czyli ocena mocy testu *post hoc* wydaje się nieprzydatna.

A zatem co jest rachowane w programach statystycznych pod kryptonimem „obserwowana moc testu" czy „moc testu *post hoc*"? Może coś, co Gillett (1994) nazywa przeciętną (*average*) mocą testu użytego w konkretnym zagadnieniu? Choć wydaje mi się, że nazwanie „mocy testu *post hoc*" „mocą przeciętną" niczego nie wyjaśnia, gdyż nie wiemy, w stosunku do czego ta przeciętna jest rachowana (oceniana). Jeśli miałaby to być przeciętna w odniesieniu do wszystkich pojedynczych hipotez alternatywnych, to jest to nierealizowalne ze względu na liczbę hipotez alternatywnych.

Poniżej przedstawiam kilka przykładów, w których obliczona została obserwowana (*post hoc*) moc testu oraz oszacowano wielkość próby w zależności od przewidywanej mocy testu. Dane pochodzą z porównania efektywności trzech terapii leczenia „łokcia tenisisty" (Kubot, 2017). Porównywane są średnie zmiennych DASH_0 (przed rozpoczęciem terapii) oraz DASH_8 (po ośmiu tygodniach od zakończenia terapii). W przykładach 3.1 i 3.2 wykorzystuję procedurę UNIANOVA

6 O merytorycznym znaczeniu obserwowanych różnic i wielkości efektu piszę w podrozdziale 4.4.

w SPSS, a nie ONEWAY, gdyż w procedurze ONEWAY nie można policzyć obserwowanej mocy testu ani wielkości efektu.

Zmienna DASH_8 – ocena funkcji bolesnej kończyny (im wyższa wartość wskaźnika, tym gorsze funkcjonowanie).

$$\text{Kod\_grupy3} = \begin{cases} 0 \text{ grupa kontrolna (terapia „klasyczna”)} \\ 1 \text{ terapia falą uderzeniową} \\ 2 \text{ terapia ultradźwiękami} \end{cases}$$

PRZYKŁAD 3.1

SPSS

```
UNIANOVA DASH_8 BY kod_grupy3
/METHOD=SSTYPE(3)
/INTERCEPT=INCLUDE
/EMMEANS=TABLES(kod_grupy3) COMPARE ADJ(BONFERRONI)
/PRINT=DESCRIPTIVE HOMOGENEITY OPOWER
/CRITERIA=ALPHA(.05)
/DESIGN=kod_grupy3.
```

Tabela 3.1. Statystyki opisowe zmiennej DASH_8 w grupach poddanych różnym terapiom (program SPSS)

kod_grupy3	Średnia	Odchylenie standardowe	*N*
0	42,136957	25,3718140	60
1	14,481873	12,8393162	30
2	32,409360	19,4763750	30
Ogółem	32,791287	24,0906995	120

Tabela 3.2. Wyniki testu Levene'a oceny równości wariancji błędu zmiennej DASH_8[a]

F	*df*1	*df*2	Istotność
10,737	2	117	,000

Test Levene'a testuje hipotezę zerową zakładającą, że wariancja błędu zmiennej zależnej jest równa we wszystkich grupach.

[a] Plan: Stała + kod_grupy3.

Tabela 3.3. Wyniki testów efektów międzyobiektowych dla zmiennej DASH_8

Źródło	Typ III sumy kwadratów	*df*	Średni kwadrat	*F*	Istotność	Parametr niecentralności	Moc obserwowana[b]
Model skorygowany	15301,907[a]	2	7650,954	16,651	,000	33,301	1,000
Stała	95112,223	1	95112,223	206,992	,000	206,992	1,000
kod_grupy3	**15301,907**	**2**	**7650,954**	**16,651**	**,000**	**33,301**	**1,000**
Błąd	53761,147	117	459,497				
Ogółem	198095,272	120					
Ogółem skorygowane	69063,055	119					

[a] *R* kwadrat = ,222 (skorygowane *R* kwadrat = ,208).
[b] Obliczone z użyciem alfa = ,05.

STATA

oneway DASH_8 kod_grupy3, tabulate

Tabela 3.4. Statystyki opisowe zmiennej DASH_8 w grupach poddanych różnym terapiom (program STATA; oczywiście średnie i odchylenia standardowe są takie same, jak w programie SPSS)

```
    |    Summary of DASH_8
kod_grupy3 |    Mean  Std. Dev.    Freq.
------------+------------------------------------
    0 | 42.136957  25.371814     60
    1 | 14.481873  12.839316     30
    2 | 32.40936  19.476375     30
------------+------------------------------------
  Total | 32.791287   24.0907     120
```

Tabela 3.5. Wyniki testów efektów „międzygrupowych" dla zmiennej DASH_8 w programie STATA

```
        Analysis of Variance
 Source      SS     df   MS      F   Prob > F
------------------------------------------------------------------------
Between groups   15301.9074   2  7650.9537   16.65   0.0000
 Within groups   53761.1471  117  459.496984
------------------------------------------------------------------------
 Total     69063.0545  119  580.361803
```

Tabela 3.6. Wyniki testu Bartletta jednorodności wariancji zmiennej DASH_8

Bartlett's test for equal variances: chi2(2) = 15.2007 Prob>chi2 = 0.001

Poniżej zamieściłem komendę programu STATA umożliwiającą oszacowanie obserwowanej mocy testu *F*-Snedecora w jednoczynnikowej ANOVA:

```
. power oneway 42.14 14.48 32.41, varerror(459.5) n1(60) n2(30) n3(30)
```

Tabela 3.7. Wyniki szacowania obserwowanej mocy testu *F*-Snedecora w jednoczynnikowej analizie wariancji dla zmiennej DASH_8, zakładając poziom istotności testu $\alpha = 0{,}05$; wymagane jest podanie wartości średnich z próby, ogólnej wartości błędu i przewidywanej liczebności w grupach

Estimated power for one-way ANOVA F test for group effect Ho: delta = 0 versus Ha: delta! = 0
Study parameters: alpha = 0.0500 N = 120 Average N = 40.0000 N1 = 60 N2 = 30 N3 = 30 delta = 0.5269 N_g = 3 m1 = 42.1400 m2 = 14.4800 m3 = 32.4100 Var_m = 127.5614 Var_e = 459.5000
Estimated power: power = 0.9996

SYSTAT

W programie tym mamy specyficzne wymagania: wariancje w grupach mają być jednakowe dla każdej z porównywanych grup (wprowadziłem odchylenie standardowe uzyskane dla całej próby) oraz liczba osób w każdej grupie ma być jednakowa. Zatem, w ogólności, oszacowanie mocy testu jest w tym przypadku wyjątkowo iluzoryczne.

```
> ANOVA
> DEPEND DASH_8
> SUBCAT KOD_GRUPY3 / EFFECT
> COVAR
> ESTIMATE / SS = TYPE3
```

Tabela 3.8. Parametry stanowiące podstawę oszacowania obserwowanej mocy testu F-Snedecora w jednoczynnikowej analizie wariancji w programie SYSTAT

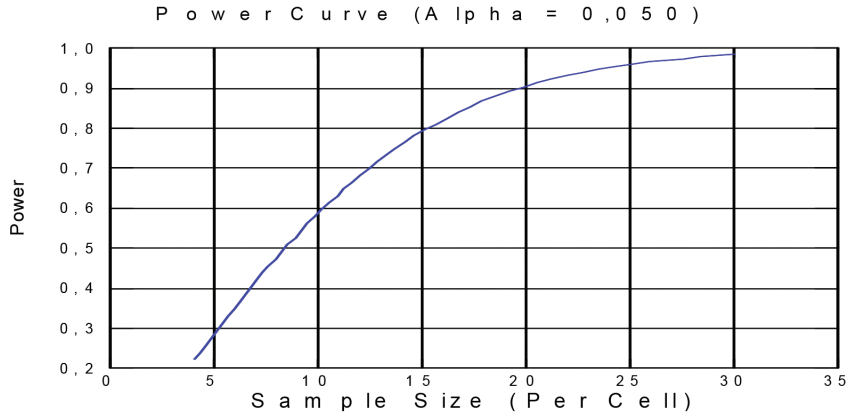
Number of Groups	:	3
Within Cell Standard Deviation	:	24,090
Mean(1)	:	42,140
Mean(2)	:	14,480
Mean(3)	:	32,410
Effect Size	:	0,476
ALPHA	:	0,050
Sample Size: Low	:	30
Sample Size: High	:	60
Increment	:	10

Non-Centrality Parameter = 0,678 * Sample Size.

Tabela 3.9. Wyniki szacowania obserwowanej mocy testu w zależności od średniej liczebności próby w komórce

Sample Size (Per Cell)	POWER
30	0,984
40	0,998
50	1,000
60	1,000

Total Sample Size = 180.



Rycina 3.1. Wykres zależności między mocą testu a wielkością próby

Powyższy wykres obrazuje moc testu (obserwowaną) w zależności od wielkości próby (w każdej z porównywanych grup) przy poziomie istotności testu $\alpha = 0{,}05$. Moc testu zależy tutaj także od wartości średniego odchylenia standardowego z próby.

Używając programów SPSS i STATA, uzyskaliśmy obserwowaną moc zastosowanego testu statystycznego, równą jedności. Konsekwencją tego będzie oszacowanie prawdopodobieństwa błędu drugiego rodzaju $\beta = 0$. Jest to informacja zbędna, trywialna, gdyż odrzuciliśmy hipotezę zerową, zatem nie mieliśmy najmniejszych szans popełnić błędu polegającego na przyjęciu fałszywej hipotezy zerowej. Podobny wynik uzyskaliśmy w programie SYSTAT, co prawda, nie dla sytuacji rzeczywistej, lecz wyimaginowanej: jednakowa liczba obserwacji w grupie i równe wariancje badanej cechy w każdej z grup.

PRZYKŁAD 3.2

W przykładzie 3.2 oprócz porachowania obserwowanej mocy testu jest jeszcze liczona wartość współczynnika η^2, jednego z mierników wielkości efektu (więcej szczegółów znajduje się w rozdziale czwartym).

SPSS

```
UNIANOVA DASH_0 BY kod_grupy3
/METHOD=SSTYPE(3)
/INTERCEPT=INCLUDE
/EMMEANS=TABLES(kod_grupy3) COMPARE ADJ(BONFERRONI)
/PRINT=ETASQ DESCRIPTIVE HOMOGENEITY OPOWER
/CRITERIA=ALPHA(.05)
/DESIGN=kod_grupy3.
```

Tabela 3.10. Statystyki opisowe zmiennej DASH_0 w grupach poddanych różnym terapiom (program SPSS)

kod_grupy3	Średnia	Odchylenie standardowe	*N*
0	53,313887	18,2438826	60
1	48,188923	16,0567165	30
2	54,049273	18,0452487	30
Ogółem	52,216493	17,6847295	120

Tabela 3.11. Wyniki testu Levene'a oceny równości wariancji błędu zmiennej DASH_0[a]

F	*df*1	*df*2	Istotność
,540	2	117	,584

Test Levene'a testuje hipotezę zerową zakładającą, że wariancja błędu zmiennej zależnej jest równa we wszystkich grupach.

[a] Plan: Stała + kod_grupy3.

Tabela 3.12. Wyniki testów efektów międzyobiektowych dla zmiennej DASH_0

Źródło	Typ III sumy kwadratów	*df*	Średni kwadrat	*F*	Istotność	Cząstkowe eta kwadrat	Parametr niecentralności	Moc obserwowana[b]
Model skorygowany	659,668[a]	2	329,834	1,056	,351	,018	2,111	,231
Stała	290357,408	1	290357,408	929,270	,000	,888	929,270	1,000
kod_grupy3	**659,668**	**2**	**329,834**	**1,056**	**,351**	**,018**	**2,111**	**,231**
Błąd	36557,541	117	312,458					
Ogółem	364404,660	120						
Ogółem skorygowane	37217,209	119						

[a] *R* kwadrat = ,018 (skorygowane *R* kwadrat = ,001).
[b] Obliczone z użyciem alfa = ,05.

STATA

oneway DASH_0 kod_grupy3, tabulate

Tabela 3.13. Statystyki opisowe zmiennej DASH_0 w grupach poddanych różnym terapiom (program STATA)

```
    |    Summary of DASH_0
kod_grupy3 |    Mean  Std. Dev.    Freq.
------------+------------------------------------
   0 | 53.313887  18.243883     60
   1 | 48.188923  16.056716     30
   2 | 54.049273  18.045249     30
------------+------------------------------------
  Total | 52.216492  17.68473     120
```

Tabela 3.14. Wyniki testów efektów „międzygrupowych" dla zmiennej DASH_0 w programie STATA

```
        Analysis of Variance
 Source      SS     df   MS      F   Prob > F
------------------------------------------------------------------------------
Between groups   659.668407   2  329.834203   1.06  0.3513
 Within groups   36557.5409  117  312.457615
------------------------------------------------------------------------------
 Total     37217.2094  119  312.749658
```

Tabela 3.15. Wyniki testu Bartletta jednorodności wariancji zmiennej DASH_0

Bartlett's test for equal variances: chi2(2) = 0.6401 Prob>chi2 = 0.726

. power oneway 53.32 48.19 54.05, vararerror(312.46) n1(60) n2(30) n3(30)

Tabela 3.16. Wyniki szacowania obserwowanej mocy testu *F*-Snedecora w jednoczynnikowej analizie wariancji dla zmiennej DASH_8, zakładając poziom istotności testu $\alpha = 0{,}05$; wymagane jest podanie wartości średnich z próby, ogólnej wartości błędu i przewidywanej liczebności w grupach

Estimated power for one-way ANOVA F test for group effect Ho: delta = 0 versus Ha: delta! = 0
Study parameters: alpha = 0.0500 N = 120 Average N = 40.0000 N1 = 60 N2 = 30 N3 = 30 delta = 0.1327 N_g = 3 m1 = 53.3200 m2 = 48.1900 m3 = 54.0500 Var_m = 5.5024 Var_e = 312.4600
Estimated power: power = 0.2313

SYSTAT

W programie tym mamy specyficzne wymagania: wariancje w grupach mają być jednakowe (wprowadziłem odchylenie standardowe dla całej próby) oraz liczba osób w każdej grupie ma być jednakowa. Zatem, w ogólności, oszacowanie mocy testu jest iluzoryczne.

> ANOVA
> DEPEND DASH_0
> SUBCAT KOD_GRUPY3 / EFFECT
> COVAR
> ESTIMATE / SS = TYPE3

Tabela 3.17. Parametry stanowiące podstawę oszacowania obserwowanej mocy testu F-Snedecora w jednoczynnikowej analizie wariancji w programie SYSTAT

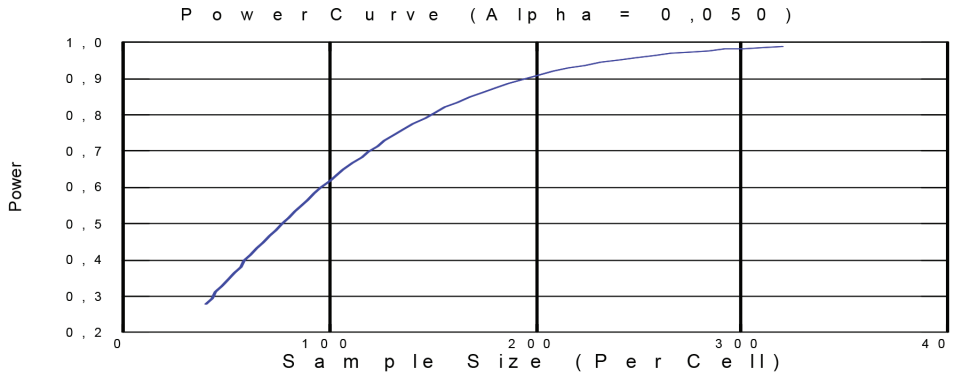
Number of Groups	:	3
Within Cell Standard Deviation	:	17,680
Mean(1)	:	53,320
Mean(2)	:	48,190
Mean(3)	:	54,050
Effect Size	:	0,147
ALPHA	:	0,050
Sample Size: Low	:	30
Sample Size: High	:	60
Increment	:	10

Non-Centrality Parameter = 0,065 * Sample Size.

Tabela 3.18. Wyniki szacowania obserwowanej mocy testu w zależności od średniej liczebności próby w komórce

Sample Size (Per Cell)	POWER
30	0,215
40	0,278
50	0,340
60	0,402

Total Sample Size = 180.



Rycina 3.2. Wykres zależności mocy testu od liczebności próby

Powyższy wykres obrazuje moc testu (obserwowaną) w zależności od wielkości próby (w każdej z porównywanych grup) przy poziomie istotności testu $\alpha = 0{,}05$. Moc testu zależy tutaj również od wartości średnich odchylenia standardowego z próby. Wielkości próby dla takiej samej mocy testu są tutaj wielokrotnie większe niż na rycinie 3.1. Jest to efekt innych wartości średnich (mniejsze zróżnicowanie) oraz odchylenia standardowego z próby.

W przykładzie 2 prawdopodobieństwo w teście zagadnienia:

$$\begin{cases} H_0: \mu_1 = \mu_2 = \mu_3 \\ H_1: \neg(\mu_1 = \mu_2 = \mu_3) \end{cases} \tag{3.10}$$

w obu programach, czyli SPSS i STATA, jest równe 0,351, a zatem jest większe od 0,05. Obserwowana moc testu została oszacowana jako 0,231, czyli odpowiadające temu prawdopodobieństwo błędu drugiego rodzaju (też tylko oszacowane) to 0,769. W czym nam pomogło oszacowanie mocy testu? Odrzucić hipotezy zerowej nie możemy, przyjąć jej także nie powinniśmy, gdyż $\beta = 0{,}769$. Gdybyśmy przyjęli hipotezę zerową, to najprawdopodobniej popełnilibyśmy błąd drugiego rodzaju.

Wyniki programu SYSTAT, mimo że tylko hipotetyczne, wskazują na podobne problemy i nie ułatwiają podjęcia racjonalnej decyzji. Zostajemy ze sformułowaniem, iż nie ma podstaw do odrzucenia hipotezy zerowej – i koniec.

Warto zwrócić uwagę na jeszcze jeden aspekt obliczania mocy testu statystycznego. Otóż, moc testu jest zależna od wielkości próby, co widać w drugim przykładzie w wynikach z programu SYSTAT. Wyraźnie widać, jak rośnie obserwowana moc testu (cokolwiek by to w rzeczywistości było) wraz ze wzrostem liczebności elementów w grupie. I moc testu oraz wielkość próby są kluczowymi składowymi w planowaniu badania statystycznego. Planując badanie, powinniśmy spróbować sobie odpowiedzieć, ilu elementów potrzebujemy w badaniu, aby osiągnąć jego cel. Zbyt mało liczna próba może uniemożliwić nam wykrycie stosunkowo słabo rozpowszechnionego zjawiska. Z kolei nakłady finansowe i np. osobowe (ankieterzy) przy zbyt licznej próbie nie zostaną zrekompensowane odpowiednim wzrostem efektywności badania – od pewnego momentu zwiększanie liczebności próby nie poprawia już efektów badania albo poprawia je bardzo nieznacznie.

Należy zwrócić uwagę na jeszcze jedno zagadnienie związane z szacowaniem mocy testu *post hoc*. Nie dla wszystkich testów statystycznych takie oszacowania są zaimplementowane w pakietach statystycznych. Nie znalazłem takich możliwości np. dla testów w modelach regresyjnych czy dla oceny niezależności dwóch zmiennych dyskretnych o liczbie kategorii większej niż dwie.

3.3. Szacowanie wielkości próby

Nie istnieje jedna, uniwersalna formuła dla oceny wielkości próby. Zarówno sposób jej wyznaczania, jak i sama wielkość próby są zależne od konkretnego modelu, poziomu istotności przewidzianego w analizie testu statystycznego, jego mocy, a bardzo często również od specyficznych właściwości badanego zjawiska. I tak, Whittemore (1981) przedstawia zagadnienie oceny wielkości próby dla regresji logistycznej, ale w sytuacji małych prawdopodobieństw odpowiedzi, czyli w sytuacji stosunkowo rzadkich zjawisk. Tablice wielkości próby w ogólnym przypadku modelowania regresji logistycznej prezentuje Hsieh (1989), zaś Lemeshow i wsp. (1988) oraz Satten i Kupper (1990) zajmują się oceną wielkości próby dla oszacowań ilorazów szans w modelach logistycznych. Schoenfeld (1983) zajmuje się oceną wielkości próby w regresyjnych modelach ryzyk proporcjonalnych (tzw. analiza przeżywalności, modele regresyjne Coxa). Connett i wsp. (1987) analizują wielkość próby i moc testu w kojarzonych badaniach typu *case-control* (jeden z typów badań epidemiologicznych), Connor (1987) – dla porównywania proporcji, zaś Lachin (1986) w bardzo specyficznym badaniu przeżywalności.

Relacje między mocą testu i liczebnością próby mogą być pomocne podczas projektowania eksperymentów, natomiast wydają się bardzo mało przydatne w przypadku badań obserwacyjnych, w których najczęściej nie mamy wpływu np. na liczebność różnych warstw w badaniu. Wyraźnie było to widać w przykładach 3.1 i 3.2 w wydrukach z programu SYSTAT. W programie założono, że warstwy są równoliczne i w rozważanej przez nas sytuacji, gdy jedna grupa liczyła 60 osób, a dwie pozostałe po 30, wyniki uzyskane w tym programie zupełnie nie przystawały do rzeczywistości.

Dla danych z przykładu 3.2 spróbuję oszacować wielkość próby.

PRZYKŁAD 3.3

SPSS

W programie SPSS nie znalazłem możliwości oszacowania wielkości próby przy zadanych wartościach poziomu istotności i mocy testu.

STATA

W programie STATA jest możliwość oszacowania liczebności próby, gdy liczebności poszczególnych grup są proporcjonalne. Analogicznie do rzeczywistego eksperymentu określiłem liczebność grupy kontrolnej jako dwa razy większą niż grup badanych.

Charakterystyka grup i wyniki jednoczynnikowej analizy wariancji dla zmiennej DASH_0 przedstawione zostały w tab. 3.13 i 3.14.

Prawdopodobieństwo w teście hipotez (3.10) jest równe 0,351, a więc większe od poziomu istotności testu $\alpha = 0{,}05$. Nie mamy podstaw do odrzucenia hipotezy

zerowej o równości wartości oczekiwanych zmiennej DASH_0 w grupach osób poddanych różnym terapiom leczenia łokcia tenisisty. Przejdziemy teraz do wyznaczenia wielkości próby dla parametrów testu $\alpha = 0{,}05$ i $\beta = 0{,}10$.

. power oneway 53.31 48.19 54.05, varerror(312.75) **power(0.9) grweights(2 1 1)**

Tabela 3.19. Parametry badania stanowiące podstawę wyznaczenia wielkości próby

alpha = 0.0500
power = 0.9000
delta = 0.1325
N_g = 3
m1 = 53.3100
m2 = 48.1900
m3 = 54.0500
Var_m = 5.4915
Var_e = 312.7500

Tabela 3.20. Oszacowane liczebności próby w poszczególnych grupach przy $\beta = 0{,}1$

Estimated sample sizes:
N = 724
Average N = 241.3333
N1 = 362
N2 = 181
N3 = 181

Przy założeniu mocy testu równej 0,9 (czyli prawdopodobieństwie błędu drugiego rodzaju równym 0,1) liczebność grupy powinna być **sześciokrotnie większa** niż była w rzeczywistości. Realistyczne wymaganie?

W tab. 3.21 przedstawione jest takie samo oszacowanie liczebności próby, ale przy założeniu $\beta = 0{,}2$.

Tabela 3.21. Oszacowane liczebności próby w poszczególnych grupach przy $\beta = 0{,}2$

Estimated sample sizes:
N = 552
Average N = 184.0000
N1 = 276
N2 = 138
N3 = 138

Przy założeniu mocy testu równej 0,8 (czyli prawdopodobieństwie błędu drugiego rodzaju równym już 0,2) liczebność grupy powinna być **ponad trzykrotnie większa**, niż była w rzeczywistości. Podejmując teraz decyzję o przyjęciu hipotezy zerowej, mamy dwukrotnie większe ryzyko popełnienia błędu drugiego rodzaju, niż w przypadku $\beta = 0{,}9$. A liczebność próby nadal jest jednak bardzo duża.

Przy grupach równolicznych uzyskujemy następujące oszacowania przy $\beta = 0{,}2$ (moc = $1 - \beta = 0{,}8$) i $\beta = 0{,}1$ (moc = $1 - \beta = 0{,}9$).

Tabela 3.22. Oszacowane liczebności próby w poszczególnych grupach przy założeniu równoliczności grup; $\beta = 0{,}2$

Estimated sample sizes:
N = 447
N per group = 149

Tabela 3.23. Oszacowane liczebności próby w poszczególnych grupach przy założeniu równoliczności grup; $\beta = 0{,}1$

Estimated sample sizes:
N = 588
N per group = 196

W obu przypadkach liczebność grupy jest wielokrotnie większa niż rzeczywista.

SYSTAT

Tabela 3.24. Wyniki szacowania liczebności próby w zależności od przyjętej mocy testu

Sample Size (Per Cell)	POWER	Total Sample Size
148	0,798	
149	**0,800**	**447**
193	0,896	
194	0,898	
195	0,900	588
196	0,901	
236	0,948	
237	0,949	
238	0,950	717
239	0,951	

Oszacowania liczebności w grupach są praktycznie takie same, jak w programie STATA i tak samo wielokrotnie większe niż rzeczywiste liczebności pacjentów w badaniu. Czy nie znając efektów badania (nie wiedząc, czy będziemy mogli odrzucić hipotezę zerową czy nie) i nie wiedząc, co oznacza moc testu (hipoteza alternatywna jest hipotezą złożoną) powinniśmy planować tak ogromne eksperymenty?

Rozdział 4. Ocena wielkości efektu

4.1. Wprowadzenie

Zagadnienie oceny wielkości efektu jest związane wyłącznie z testowaniem hipotez według teorii Neymana-Pearsona. Przy tej ocenie wykorzystywane są składowe testowania hipotez, np. wartość statystyki chi-kwadrat w teście oceny niezależności dwóch zmiennych dyskretnych, sumy kwadratów z analizy wariancji czy wartości R^2 z testów wykorzystywanych w regresji liniowej.

Oto króciutkie przypomnienie pewnych teoretycznych problemów pojawiających się podczas testowania hipotez statystycznych według teorii Neymana-Pearsona. Hipoteza zerowa jest hipotezą prostą, sformułowaną w postaci, którą wolimy odrzucić, np. H_0: $\mu_1 = \mu_2$, H_0: $\rho = 0$ itp. Warto zwrócić uwagę na pewien aspekt zagadnienia oceny wielkości efektu. Skoro w tej ocenie wykorzystywane są wartości statystyk będących podstawą testów, to ich wielkości będą bezpośrednio wpływały na wielkość miernika efektu. W konsekwencji im większa wartość statystyki, tym większy miernik efektu. Z drugiej strony, wartość statystyki będącej podstawą testu jest wykorzystywana do obliczenia prawdopodobieństwa, które z kolei porównywane jest z poziomem istotności testu i decyzja dotycząca porachowanego prawdopodobieństwa jest dwuwartościowa: $p < \alpha$, $p > \alpha$. Czy zatem testowanie hipotez statystycznych i ocena wielkości efektu są kompatybilnymi sposobami oceny zależności? Gdy odrzucamy hipotezę zerową ($p < \alpha$), to wartość miernika wielkości efektu możemy interpretować jako ocenę siły zależności. Ale taka interpretacja oznacza najpierw porównywanie wartości statystyki będącej podstawą testu, a następnie porównywanie prawdopodobieństw odpowiadających tym wartościom. A tego, na podstawie teorii testowania hipotez Neymana-Pearsona, nie mamy prawa robić. Wydaje mi się, że powstał klincz, który bez zmiany paradygmatu w testowaniu hipotez statystycznych i ujednolicenia teorii jest nierozwiązywalny.

Aby zakończyć te przygnębiające rozważania bardziej optymistycznym akcentem, zacytuję fragment z przedmowy Lehmanna (1968).

> […] Znacznie ważniejszy jest fakt, że wiele zagadnień, które tradycyjnie były formułowane w terminach testowania hipotez, są w rzeczywistości problemami wielodecyzyjnymi, w których w przypadku odrzucenia testowanej hipotezy ma się do wyboru kilka decyzji. Znalezienie odpowiednich rozwiązań dla tego typu problemów jest obecnie najważniejszym zadaniem statystyki i zajmuje wiele miejsca w bieżącej literaturze. […] W zasadzie wydaje się rzeczą prawdopodobną, że te tradycyjne testy pozostaną użyteczne ze względu na swą prostotę, nawet wtedy, gdy będziemy już mieli bardziej kompletną teorię postępowań wielodecyzyjnych.

Możemy więc spróbować potraktować ocenę wielkości efektu przy odrzuceniu hipotezy zerowej w teorii Neymana-Pearsona jako prymitywną i nieudokumentowaną teoretycznie próbę rozwiązania pewnego, też nie do końca uświadomionego, problemu wielodecyzyjnego.

4.2. Ocena wielkości efektu

Tytułem wprowadzenia dwa cytaty z książki A. Fielda *Discovering Statistics Using SPSS*. Cytaty te ilustrują niefrasobliwość i nieodpowiedzialność w używaniu terminologii statystycznej, która w konsekwencji prowadzi do omawianego wcześniej pomieszania teorii wnioskowania statystycznego. „Proponowano wiele miar jako wielkość efektu, lecz najbardziej znane spośród nich to współczynnik *d* Cohena, współczynnik korelacji Pearsona i iloraz szans”[1] (Field, 2009, rozdz. 2, s. 52).

> 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; 1992a; 1992b) zaproponował szeroko wykorzystywane interpretacje wielkości *r*, które mogą świadczyć o dużym albo małym efekcie:
>
> *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[2]
>
> (Field, 2009, rozdz. 2, s. 57).

1 „Many measures of effect size have been proposed, the most common of which are Cohen's d, Pearson's correlation coefficient r and the odds ratio”.

2 „Effect sizes are useful because they provide an objective measure of the importance of an effect. So, it doesn't matter what effect you're looking for, what variables have been measured, or how those variables have been measured – we know that a correlation coefficient of 0 means there is no effect, and a value of 1 means that there is a perfect effect. Cohen (1988, 1992) has also made some widely used suggestions about what constitutes a large or small effect:

r = .10 (small effect): in this case the effect explains 1% of the total variance,

r = .30 (medium effect): the effect accounts for 9% of the total variance,

r = .50 (large effect): the effect accounts for 25% of the variance”.

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, nieodpowiedzialność w działaniu. Można by to pominąć milczeniem, gdyby nie mieszało w głowach osób dopiero uczących się statystyki. Aby uzyskać wiarygodne wyniki jakiegokolwiek badania, musimy najpierw dokładnie określić jego cel, sposób realizacji, a więc sprecyzować, co będziemy mierzyli albo klasyfikowali, według jakich kryteriów. To prowadzi do wskazania, z jakiego typu zmiennymi będziemy mieli do czynienia, a cel badania pozwoli określić, jakich zależności między zmiennymi będziemy poszukiwali i jakimi metodami będziemy to robili.

Nieprawdą jest również, że wielkości efektu stanowią obiektywną miarę ważności efektu. Po pierwsze, czy „obiektywna" oznacza, że jest to miara w jakikolwiek sposób prawdziwa? Że możemy tę miarę „przyłożyć do rzeczywistości", „[…] że współczynnik korelacji równy 0 oznacza brak efektu, a jego wartość równa 1 oznacza, że efekt jest pełny (kompletny)"? Cóż oznacza stwierdzenie, że efekt jest pełny? Sformułowania te świadczą, iż autor nie ma pojęcia o zagadnieniach teorii miary (a prawdopodobieństwo jest unormowaną miarą) i nie słyszał nic o zbiorach miary zero, które nie są zbiorami pustymi i mogą wpływać na zachodzenie albo nie odpowiednich zdarzeń.

> 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 rozumiany, często używany i tak, przyznaję się, naprawdę go lubię! Przekształcając wartość *t* w wartość współczynnika korelacji *r*, co jest naprawdę łatwe, możemy użyć następującego równania[3] (np. Rosenthal, 1991; Rosnow, Rosenthal, 2005) (Field, 2009, rozdz. 9, s. 332):

$$r = \sqrt{\frac{t^2}{t^2+df}} \tag{4.1}$$

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 r. Przy założeniu prawdziwości hipotezy $H_0 : \rho = 0$ statystyka

$$t = \frac{r}{\sqrt{1-r^2}}\sqrt{n-2} \tag{4.2}$$

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

3 „Even though our *t*-statistics is statistically significant, this doesn't mean our effect is important in practical terms. To discover whether the effect is substantive we need to use what we know about effect sizes. I'm going to stick with the effect size *r* because it's widely understood, frequently used, and yes, I'll admit it, I actually like it! Converting a *t*-value into an *r*-value, is actually really easy; we can use the following equation".

Przekształćmy wzór (4.2) we wzór (4.1). Dla uproszczenia zapisu oznaczmy $n - 2 = df$.

$$t = \frac{r}{\sqrt{1-r^2}}\sqrt{df} \equiv t^2 = \frac{r^2 \cdot df}{1-r^2} \equiv t^2(1-r^2) = r^2 \cdot df \equiv t^2 - t^2 \cdot r^2 = r^2 \cdot df$$
$$\equiv t^2 = t^2 \cdot r^2 + r^2 \cdot df \equiv t^2 = (t^2 + df) \cdot r^2 \equiv r^2 = \frac{t^2}{t^2+df} \equiv r = \sqrt{\frac{t^2}{t^2+df}} \quad (4.3)$$

Otrzymanie wzoru (4.1) jest, jak widać, 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, żadna miara czy jakikolwiek wynik statystyczny nie zwalnia badacza z obowiązku myślenia – myślenia w kategoriach merytorycznych. Jak już podkreślałem wcześniej, statystyka pełni rolę służebną, tylko pomocniczą w realizacji badań.

Gigerenzer i Marewski (2015) zauważają:

> Jeśli statystycy zgodziliby się na jedną rzecz, byłoby to uznanie, iż wnioskowanie statystyczne nie może być robione mechanicznie. Mimo zjadliwych sporów w innych kwestiach, Ronald Fisher i Jerzy Neyman, dwóch najbardziej wpływowych statystyków XX w., w tej sprawie mówili jednym głosem. Dobra nauka wymaga zarówno narzędzi statystycznych, jak i przemyślanej oceny, w jakim celu konstruowany jest model, jakie hipotezy są testowane i jakich narzędzi należy użyć. Praktykujący statystycy funkcjonują w oparciu o „statystyczną skrzynkę narzędziową" i wykorzystują swoje kompetencje przy wyborze odpowiedniego narzędzia; naukowcy nauk społecznych, w przeciwieństwie do nich, skłaniają się do wykorzystywania pojedynczego narzędzia.

Jak ujął to psycholog Abraham Maslow (1966): „jeśli wszystkim, czym dysponujemy, jest młotek, każda rzecz wygląda jak gwóźdź".

Wielkość efektu (*effect size*) bywa także nazywana siłą zależności (*strength of association*, *treatment magnitude*). 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 wiąże

się z odrzuceniem hipotezy bądź zaakceptowaniem hipotezy. Od tego czasu zaproponowano wiele różnych miar oceny wielkości efektu. W tym rozdziale nie będę przedstawiał ich wszystkich, lecz ograniczę się do tych, które są zaimplementowane w programach statystycznych, np. SPSS czy STATA, a także są stosunkowo łatwe do policzenia na podstawie danych z wyników analizy, jak np. *d* Cohena.

4.3. Mierniki oceny wielkości efektu

4.3.1. Dwie najprostsze sytuacje analizy danych

Do dwóch najprostszych sytuacji, w których wykorzystujemy metody statystyczne zaliczam porównywanie dwóch wartości oczekiwanych i ocenę niezależności miedzy dwiema zmiennymi dyskretnymi.

4.3.1.1. Porównywanie dwóch wartości oczekiwanych

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

$$d = \frac{\mu_1 - \mu_2}{\sigma} \tag{4.4}$$

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:

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s_{wspólne}}; \quad s_{wspólne} = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2}} \tag{4.5}$$

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s_{wspólne}}; \quad s_{wspólne} = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}} \tag{4.6}$$

gdzie n_1 i n_2 to liczebności próbek, na podstawie których obliczane były średnie i wariancje z próby. Wzór (4.5) pochodzi z opracowania Thalheimera i Cooka (2002), zaś wzór (4.6) z pracy Volkera (2006).

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

Podawanie izolowanych punktów dla określania siły efektu nie ma sensu. Konieczne jest przedstawienie przedziałów liczbowych (abstrahując od faktu, że jest to pewna „prawda objawiona", czyli, jak się komuś wydawało, że „będzie dobrze"). Ale przyjmijmy tę prawdę, określając odpowiednie przedziały:
$d < 0{,}2$ – brak efektu,
$0{,}2 \leq d < 0{,}5$ – efekt słaby,
$0{,}5 \leq d < 0{,}8$ – efekt średni,
$d \geq 0{,}8$ – efekt silny.

Tak arbitralnie przyjmowane granice dla tego parametru czy innych nie mają wiele wspólnego z merytorycznym znaczeniem obserwowanych efektów – jest to ciągle pewna zabawa na liczbach i próba zwolnienia się z głębokiego przemyślenia uzyskanych rezultatów, zarówno w terminach statystycznych, jak i merytorycznych.

PRZYKŁAD 4.1

Porównajmy wartości oczekiwane cholesterolu całkowitego (zmienna: cholest), frakcji HDL cholesterolu (zmienna: HDL), poziomu cukru na czczo (zmienna: cukier) oraz wyników pierwszego pomiaru ciśnienia skurczowego (zmienna: skurcz1) w grupach określonych przez wartości zmiennej ukl_kraz (dane: Dudek, 2007). Zmienna ukl_kraz jest zmienną dwustanową:

$$\text{ukl\_kraz} = \begin{cases} 0 & \text{nie zdiagnozowano chorób układu krążenia} \\ 1 & \text{zdiagnozowano jakąś chorobę układu krążenia} \end{cases}$$

```
oneway cholest hdl cukier skurcz1 by ukl_kraz
/statistics descriptives homogeneity brownforsythe welch
/missing analysis.
```

Tabela 4.1. Statystyki opisowe analizowanych zmiennych w grupach wyznaczonych przez wartości zmiennej ukl_kraz (program SPSS)

Statystyki opisowe									
Zmienne		*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
skurcz1	,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

Tabela 4.2. Wyniki testu Levene'a oceny równości wariancji analizowanych zmiennych

Zmienne	Test Levene'a	*df*1	*df*2	Istotność
cholest	,011	1	436	,915
HDL	,073	1	436	,787
cukier	4,399	1	436	,037
skurcz1	80,291	1	435	,000

Tabela 4.3. Wyniki porównywania wartości oczekiwanych zmiennych „cholest" i „HDL" testem *F*-Snedecora

Zmienne		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			

Tabela 4.4. Wyniki porównywania wartości oczekiwanych zmiennych „cukier" i „skurcz1" odpornymi testami Welcha i Browna-Forsythe'a

		Statystyka[a]	*df*1	*df*2	Istotność
cukier	Welch	3,075	1	222,715	,081
	Brown-Forsythe	3,075	1	222,715	,081
skurcz1	Welch	57,880	1	165,728	,000
	Brown-Forsythe	57,880	1	165,728	,000

[a] Rozkład *F* asymptotyczny.

Dla dwóch badanych cech – cholesterol całkowity („cholest") i HDL – nie możemy odrzucić hipotez o jednorodności wariancji, natomiast dla zmiennych „cukier" i ciśnienie skurczowe („skurcz1") wariancje w porównywanych grupach są różne. Wynikają z tego oczywiste i znane konsekwencje: dla zmiennych cholesterol całkowity i HDL zastosujemy klasyczny test *F*-Snedecora do porównywania wartości oczekiwanych, zaś dla zmiennych cukier i ciśnienie skurczowe – testy Welcha i Browna-Forsythe'a, które są odporne (a nie mocne, moc testu to coś zupełnie innego) na niespełnianie założenia jednorodności wariancji badanej cechy w porównywanych grupach.

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 badanej zmiennej w porównywanych grupach.

W procedurze jednoczynnikowa ANOVA w SPSS-ie nie ma możliwości obliczenia wielkości efektu. Ponieważ w programie obliczane są odpowiednie sumy kwadratów, można to zrobić „ręcznie".

Tabela 4.5. Obliczenie mierników wielkości efektu *d* Cohena i η^2 oraz obserwowanej mocy testu dla zmiennych analizowanych w przykładzie 4.1

Badana zmienna	ukl_kraz	Średnia	SD	$\bar{x}_1 - \bar{x}_2$	$s_{wspólne}$	*d* Cohena*	Cząstkowe η^2	Obserwowana moc testu
Cholest	0	199,87	36,40	–13,00	36,23	–0,36	0,027	0,933
	1	212,87	35,84					
HDL	0	61,55	13,00	3,08	12,86	0,24	0,012	0,636
	1	58,47	12,53					
Cukier	0	84,92	11,27	–2,31	11,94	–0,19	0,0008	0,463
	1	87,23	13,34					
Ciśnienie skurczowe (skurcz1)	0	129,50	9,40	–12,13	12,31	–0,99	0,171	1,000
	1	141,63	17,24					

* Obliczone według wzoru (4.6).

Znak miernika *d* Cohena nie jest ważny dla wielkości efektu – jest on zależny od tego, którą średnią od której odejmujemy, więc nie będę się nim zajmował. Ograniczę się do dyskusji mierników dla dwóch zmiennych, a mianowicie HDL i cukru we krwi. Wyniki testu hipotezy o równości wartości oczekiwanych dla zmiennej HDL doprowadziły do podjęcia decyzji, że wartości oczekiwane w grupach są różne (mamy tylko dwie porównywane grupy, więc niepotrzebne są testy porównań wielokrotnych). Decyzja ta dotyczy wartości oczekiwanych, czyli parametrów teoretycznych, traktowanych niekiedy jako prawdziwe wartości tych parametrów. Decyzja wynikająca z testowania hipotez statystycznych zawsze 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. Jednak testowanie hipotez i obliczanie wielkości efektów odbywają się w różnych przestrzeniach. Ile mają ze sobą wspólnego? Nie należy również 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 analizy statystycznej nie jest argumentem rozstrzygającym, lecz 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 istnieje) ułatwia nam odnalezienie się w tej sytuacji? Sądzę, że nie. Ponadto, zwróćmy uwagę na wartość obserwowanej mocy testu. Czy wartość mocy testu równa 0,463 pozwala na przyjęcie hipotezy zerowej, iż wartości oczekiwane się nie różnią w obu grupach? Zatem zgodzimy się, że te wartości oczekiwane są różne, ale zależność jest słaba. Może zatem warto całkowicie zrezygnować z testowania hipotez, co sugerują niektórzy badacze? 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, np. lekarzowi? Szerzej o merytorycznej wartości obserwowanych różnic piszę w podrozdziale 4.4.

W tab. 4.1 występuje jeszcze jeden miernik wielkości efektu – cząstkowe η^2. Szerzej miernik ten jest omawiany w podrozdziale 4.3.3. Dla zmiennej „HDL" przy $\eta_p^2 = 0{,}012$ i $p = 0{,}021$ w teście, w oparciu o przedstawione wyżej wytyczne, podejmujemy decyzję, iż średnie się różnią, ale zależność jest słaba. W przypadku zmiennej „cukier", dla której $p = 0{,}081$ w teście i $\eta_p^2 = 0{,}0008$, moglibyśmy powiedzieć, że nie ma zależności, gdyby nie moc testu równa tylko 0,463.

4.3.1.2. Ocena niezależności dwóch zmiennych dyskretnych

Podstawowymi testami statystycznymi do oceny niezależności dwóch zmiennych dyskretnych są: test chi-kwadrat niezależności, nazywany 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 (w SPSS-ie tabelą krzyżową). Efektem testowania hipotez:

$$\begin{cases} H_0\text{: zmienne } X \text{ i } Y \text{ są niezależne} \\ H_1\text{: zmienne } X \text{ i } Y \text{ nie są niezależne} \end{cases} \tag{4.7}$$

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), to mamy do dyspozycji całą gamę różnego rodzaju mierników siły zależności między badanymi zmiennymi. Nie będę ich tutaj wymieniał – odsyłam Czytelnika do podręcznika (Szymczak, 2018).

Cohen (1988) zaproponował jako miernik wielkości efektu:

$$\Phi = \sqrt{\frac{\chi^2}{n}} = w \tag{4.8}$$

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 badanych elementów 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. Znów mamy tu do czynienia z izolowanymi wartościami, a nie przedziałami.

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 najlepszą we wszystkich sytuacjach. Dla tablic $r \times c$ (tablica o r wierszach i c kolumnach) 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? Można sformułować jeszcze trzecie zastrzeżenie związane ze skalą, na jakiej mierzone są zmienne dyskretne. Wszak używamy innych mierników siły zależności w przypadku zmiennych nominalnych, zaś innych w przypadku zmiennych porządkowych.

4.3.1.3. Dokładny test Fishera (Woolson, 1987)

Problemy pojawiające się w związku ze stosowaniem wielkości efektów dla tabel kontyngencji, w przypadku konieczności zastosowania dokładnego testu Fishera przedstawię dla tabeli 2×2 jako najbardziej intuicyjnej. Warunkiem stosowalności testu chi-kwadrat niezależności jest, aby liczebności oczekiwane w każdej komórce tablicy kontyngencji były nie mniejsze od 5. Gdy chociaż w jednej komórce tabeli liczebność oczekiwana jest mniejsza od 5, to do oceny niezależności dwóch zmiennych dyskretnych powinniśmy zastosować dokładny test Fishera. Poniżej przedstawię procedurę realizacji dokładnego testu Fishera, nie wchodząc w szczegóły teoretyczne.

Tabela 4.6. Najprostsza tabela krzyżowa 2 × 2 (dane wyjściowe)

Zmienna *X*	Zmienna *Y*		Suma w wierszach
	1	2	
1	*a*	*c*	*a* + *c*
2	*b*	*d*	*b* + *d*
Suma w kolumnach	*a* + *b*	*c* + *d*	*a* + *b* + *c* + *d* = *n*

Dla tych danych obliczane jest prawdopodobieństwo:

$$p_a = \left[\frac{(a+b+c+d)!a!b!c!d!}{(a+c)!(b+d)!(a+b)!(c+d)!}\right]^{-1} \tag{4.9}$$
$$n! = n \cdot (n-1) \cdot (n-2) \cdot \cdots \cdot 2 \cdot 1 \; 0! = 1 \;\; 1! = 1$$

W następnym kroku tworzona jest tabela:

Tabela 4.7. Tabela krzyżowa 2 × 2 po pierwszym kroku obliczania prawdopodobieństwa w dokładnym teście Fishera

Zmienna *X*	Zmienna *Y*		Suma w wierszach
	1	2	
1	*a* – 1	*c* + 1	*a* + *c*
2	*b* + 1	*d* – 1	*b* + *d*
Suma w kolumnach	*a* + *b*	*c* + *d*	*a* + *b* + *c* + *d* = *n*

i obliczane prawdopodobieństwo:

$$p_{a-1} = \left[\frac{(a+b+c+d)!(a-1)!(b+1)!(c+1)!(d-1)!}{(a+c)!(b+d)!(a+b)!(c+d)!}\right]^{-1} \tag{4.10}$$

Znów odejmujemy 1 od *a* – 1 i *d* – 1, a dodajemy 1 do *b* + 1 i *c* + 1 oraz obliczamy prawdopodobieństwo p_{a-2}. Proces kontynuujemy tak długo, aż liczebność którejś z komórek na głównej przekątnej będzie równa 0. Decyzję dotyczącą hipotez podejmujemy, porównując sumę prawdopodobieństw: $p_a + p_{a-1} + \ldots$ z poziomem istotności testu. Jest oczywiste, że obliczenia w teście dokładnym Fishera są bardzo uciążliwe do przeprowadzenia.

Dla dużych *n* możemy jednak użyć przybliżonego rozkładu chi-kwadrat. Można udowodnić, iż statystyka:

$$\frac{n \cdot [|a \cdot d - b \cdot c| - n/2]^2}{(a+b) \cdot (c+d) \cdot (a+c) \cdot (b+d)} \tag{4.11}$$

ma przybliżony rozkład chi-kwadrat z jednym stopniem swobody. By można było użyć tego przybliżenia, *n* powinno być dostatecznie duże, tak aby:

$$\frac{a+b}{n} \cdot (a+c) \geq 5,\ \frac{a+b}{n} \cdot (b+d) \geq 5,\ \frac{c+d}{n} \cdot (a+c) \geq 5, \frac{c+d}{n} \cdot (b+d) \geq 5 \tag{4.12}$$

Z drugiej strony, przywołajmy cytat z książki Magiery (2007): „dokładny test Fishera stosowany jest do sprawdzenia hipotezy o niezależności lub hipotezy o jednakowych rozkładach zmiennych losowych *X* i *Y* w przypadku, gdy dane dostępne są w formie tablicy wielodzielczej 2 × 2, przy czym częstości $n_{i,j}$, gdzie $i, j = 1, 2$, są małe i całkowita liczba obserwacji jest mniejsza od 20. Z powodu małej liczby danych, zamiast granicznego rozkładu χ^2 stosowanego w odpowiednich modelach dotyczących testu chi-kwadrat niezależności oraz testu chi-kwadrat jednorodności, w teście tym wykorzystuje się rozkład dokładny, obliczając wartości prawdopodobieństw otrzymania określonego układu zaobserwowanych częstości zgodnie z rozkładem hipergeometrycznym".

Gwoli przypomnienia: hipotezę zerową o niezależności dwóch zmiennych dyskretnych możemy sformułować np. w postaci jednakowych rozkładów częstości wartości zmiennej *Y* dla różnych wartości zmiennej *X*, a hipotezę alternatywną o ich zależności jako zróżnicowanie analogicznych rozkładów częstości.

Oryginalny dokładny test Fishera został uogólniony na tablice kontyngencji większe niż 2 × 2 (Freeman, Halton, 1951), zaś algorytmy jego obliczania opracowane zostały przez Mehtę i Patela (1983, 1986).

4.3.1.4. Przykłady

PRZYKŁAD 4.2

Oceńmy niezależność dwóch zmiennych dyskretnych: grupa (grupa zawodowa) i palenie tytoniu. Kategorie zmiennej grupa: 1 → strażacy, 2 → pracownicy służb więziennych, 3 → policjanci. Kategorie zmiennej palenie tytoniu (nazwa zmiennej: palenie3): 0 → osoby, które nie palą i nie paliły w przeszłości, 1 → osoby aktualnie palące, 2 → byli palacze (dane: Dudek, 2007).

```
CROSSTABS
/TABLES=palenie3 BY grupa
/FORMAT=AVALUE TABLES
/STATISTICS=CHISQ
/CELLS=COUNT COLUMN
/COUNT ROUND CELL
/METHOD=EXACT TIMER(5).
```

Tabela 4.8. Tabela krzyżowa dla zmiennych „grupa" i „palenie3"

			Grupa			Ogółem
			1,0	2,0	3,0	
palenie3	,0	Liczebność	37	33	81	151
		% z „grupa"	36,6%	37,1%	33,1%	34,7%
	1,0	Liczebność	44	39	99	182
		% z „grupa"	43,6%	43,8%	40,4%	41,8%
	2,0	Liczebność	20	17	65	102
		% z „grupa"	19,8%	19,1%	26,5%	23,4%
Ogółem		Liczebność	101	89	245	435
		% z „grupa"	100,0%	100,0%	100,0%	100,0%

Tabela 4.9. Wyniki testu niezależności zmiennych „grupa" i „palenie3"

	Wartość	*df*	Istotność asymptotyczna (dwustronna)
Chi-kwadrat Pearsona	3,001[a]	4	,558
Iloraz wiarygodności	3,039	4	,551
Test związku liniowego	1,743	1	,187
N ważnych obserwacji	435		

[a] 0,0% komórek (0) ma liczebność oczekiwaną mniejszą niż 5; minimalna liczebność oczekiwana wynosi 20,87.

Tabela 4.10. Symetryczne mierniki siły zależności między zmiennymi „grupa" i „palenie3"

		Wartość	Istotność przybliżona
Nominalna przez Nominalna	Phi	,083	,558
	V Kramera	,059	,558
	Współczynnik kontyngencji	,083	,558
N ważnych obserwacji		435	

Ponieważ zmienna „grupa" jest zmienną nominalną, użyłem mierników symetrycznych dla zmiennych nominalnych. Wartość miernika $\Phi = w$ jest równa 0,083, co oznacza praktyczny brak zależności między tymi dwiema zmiennymi. I w realiach tego przykładu wszystko się zgadza: nie mamy podstaw do odrzucenia hipotezy zerowej o niezależności zmiennych ($p = 0{,}558$), we wszystkich komórkach tablicy kontyngencji liczebności oczekiwane są większe od 5, a więc można stosować test chi-kwadrat Pearsona, zaś miernik wielkości efektu jest równy 0,083, co wskazuje na brak zależności, czyli możemy zgodzić się, iż zmienne „grupa" i „palenie3" są zmiennymi niezależnymi. Można nawet uznać, że obliczenie wielkości efektu pozwala nam podjąć decyzję o niezależności zmiennych (w wyniku testowania hipotez jedynie nie możemy odrzucić hipotezy zerowej o niezależności).

PRZYKŁAD 4.3

Oceńmy niezależność (albo zależność) zmiennych grupa_wiekowa i ukl_kraz (układ krążenia). Zmienna „ukl_kraz" jest zmienną dwustanową:

$$\text{ukl\_kraz} = \begin{cases} 0 & \text{nie zdiagnozowano chorób układu krążenia,} \\ 1 & \text{zdiagnozowano jakąś chorobę układu krążenia} \end{cases}$$

zaś zmienna „grupa wiekowa" to zmienna trzystanowa:

$$\text{grupa\_wiekowa} = \begin{cases} 1 & \text{wiek do 35 lat (domknięty),} \\ 2 & \text{wiek od 35 lat (otwarty) do 45 lat (domknięty),} \\ 3 & \text{wiek powyżej 45 lat} \end{cases}$$

gdzie słowo „domknięty" oznacza, że wiek 35 lat i 45 lat należy do określonego przedziału, zaś otwarty, że 35 lat już nie należy do tego przedziału.

```
CROSSTABS
/TABLES=ukl_kraz BY grupa_wieku
/FORMAT=AVALUE TABLES
/STATISTICS=CHISQ
/CELLS=COUNT COLUMN
/COUNT ROUND CELL
/METHOD=EXACT TIMER(5).
```

Tabela 4.11. Tabela krzyżowa dla zmiennych „grupa_wiekowa" i „ukl_kraz"

			grupa_wiekowa			Ogółem
			1,00	2,00	3,00	
ukl_kraz	,0	Liczebność	151	124	29	304
		% z grupa_wiekowa	81,2%	64,2%	47,5%	69,1%
	1,0	Liczebność	35	69	32	136
		% z grupa_wiekowa	18,8%	35,8%	52,5%	30,9%
Ogółem		Liczebność	186	193	61	440
		% z grupa_wiekowa	100,0%	100,0%	100,0%	100,0%

Tabela 4.12. Wyniki testu niezależności zmiennych „grupa_wiekowa" i „ukl_kraz"

	Wartość	*df*	Istotność asymptotyczna (dwustronna)
Chi-kwadrat Pearsona	28,119[a]	2	,000
Iloraz wiarygodności	28,204	2	,000
Test związku liniowego	28,055	1	,000
N ważnych obserwacji	440		

[a] 0,0% komórek (0) ma liczebność oczekiwaną mniejszą niż 5; minimalna liczebność oczekiwana wynosi 18,85.

Tabela 4.13. Kierunkowe mierniki siły zależności między zmiennymi „grupa_wiekowa" i „ukl_kraz"

			Wartość	Asymptotyczny błąd standardowy[a]	Przybliżone T[b]	Istotność przybliżona
Nominalna przez Nominalna	Lambda	symetryczna	,078	,046	1,642	,101
		zmienna zależna: ukl_kraz	,022	,057	,384	,701
		zmienna zależna: grupa_wiekowa	,109	,063	1,633	,102
	Tau Goodmana i Kruskala	zmienna zależna: ukl_kraz	,064	,023		,000[c]
		zmienna zależna: grupa_wiekowa	,030	,012		,000[c]
	Współczynnik niepewności	symetryczna	,040	,015	2,703	,000[d]
		zmienna zależna: ukl_kraz	,052	,019	2,703	,000[d]
		zmienna zależna: grupa_wiekowa	,032	,012	2,703	,000[d]
Porządkowa przez Porządkowa	*d* Somersa	symetryczna	,236	,043	5,391	,000
		zmienna zależna: ukl_kraz	,201	,037	5,391	,000
		zmienna zależna: grupa_wiekowa	,287	,052	5,391	,000

[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.

Tabela 4.14. Symetryczne mierniki siły zależności między zmiennymi „grupa_wiekowa" i „ukl_kraz"

		Wartość	Asymptotyczny błąd standardowy[a]	Przybliżone T[b]	Istotność przybliżona
Nominalna przez Nominalna	Phi	,253			,000
	V Kramera	,253			,000
	Współczynnik kontyngencji	,245			,000
Porządkowa przez Porządkowa	Tau-b Kendalla	,240	,044	5,391	,000
	Tau-c Kendalla	,245	,045	5,391	,000
	Gamma	,446	,074	5,391	,000
N ważnych obserwacji		440			

[a] Nie zakładając hipotezy zerowej.
[b] Użyto asymptotycznego błędu standardowego przy założeniu hipotezy zerowej.

Obie zmienne dyskretne potraktowałbym jako zmienne mierzone na skali porządkowej. Prawdopodobieństwo w teście chi-kwadrat Pearsona jest mniejsze od 0,0005, więc tym samym odrzucamy hipotezę zerową i przyjmujemy hipotezę alternatywną: zmienne nie są niezależne. Wartość $\Phi = w = 0{,}253$, czyli zależność według Cohena jest zależnością słabą. Do testowania zagadnienia niezależności/zależności tych dwóch zmiennych mogliśmy użyć testu chi-kwadrat niezależności, gdyż w żadnej z komórek tabeli kontyngencji liczebność oczekiwana nie była mniejsza od 5. Jedyny, ale kluczowy problem, to użycie jako miernika wielkości efektu parametru Φ, który mierzy siłę zależności między dwiema zmiennymi nominalnymi. A np. parametr gamma, wykorzystywany w przypadku zmiennych porządkowych, jest równy 0,446, czyli jest prawie dwa razy większy od Φ. W propozycjach Cohena zabrakło rozwiązania dla zmiennych porządkowych. Dlaczego jednak mamy traktować zmienne porządkowe jako nominalne? Może lepiej nie używać miernika wielkości efektu?

PRZYKŁAD 4.4

Rozważmy teraz pewną hipotetyczną sytuację w grupie kobiet (w badanej próbie jest ich tylko 24). Oceńmy zależność między zmienną 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.

```
CROSSTABS
/TABLES=hobby BY grupa
/FORMAT=AVALUE TABLES
/STATISTICS=CHISQ
/CELLS=COUNT COLUMN
/COUNT ROUND CELL
/METHOD=EXACT TIMER(5).
```

Tabela 4.15. Tabela krzyżowa dla zmiennych „grupa" i „hobby"

			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%

Tabela 4.16. Wyniki testu niezależności zmiennych „grupa" i „hobby"

	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.

Tabela 4.17. Kierunkowe mierniki siły zależności między zmiennymi „grupa" i „hobby"

			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.

Tabela 4.18. Symetryczne mierniki siły zależności między zmiennymi „grupa" i „hobby"

		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		

Pod tabelą 4.16 znajduje się informacja, że „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 jest właśnie na podstawie wartości statystyki chi-kwadrat i jest równa 0,399:

$$\Phi = w = \sqrt{\frac{3{,}830}{24}} = 0{,}399 \tag{4.13}$$

Wartość statystyki będącej podstawą dokładnego testu Fishera ma w SPSS-ie inną wartość i analogiczne obliczenia, jak dla chi-kwadrat, dadzą inny wynik:

$$\sqrt{\frac{3{,}272}{24}} = 0{,}369 \tag{4.14}$$

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 stosowaniem testu. Rodzą się tu jeszcze inne pytania, np. czy wartość statystyki w dokładnym teście Fishera to wartość asymptotycznego rozkładu chi-kwadrat, czy mamy prawo używać przybliżonego rozkładu chi-kwadrat?

Przypomnijmy podstawy testu chi-kwadrat niezależności (Greń, 1968): „Populacja generalna jest równocześnie badana ze względu na dwie cechy, niekoniecznie mierzalne. Z populacji tej wylosowano niezależnie dużą próbę o liczebności *n* elementów. Wyniki próby klasyfikujemy w tablicę o *r* wierszach i *s* kolumnach". Podstawą testu niezależności chi-kwadrat (Pearsona) jest statystyka:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{\left(n_{ij} - np_{ij}\right)^2}{np_{ij}} \tag{4.15}$$

gdzie:

n_{ij} – liczebności obserwowane w komórkach,

np_{ij} – liczebności oczekiwane w komórkach.

Statystyka (4.17) ma przy założeniu prawdziwości hipotezy H_0 o niezależności cech asymptotyczny rozkład χ^2 z $(r - 1) \times (s - 1)$ stopniami swobody.

Jak to wygląda w przypadku dokładnego testu Fishera? W programie SPSS w większości sytuacji liczona jest jakaś wartość statystyki (w instrukcji SPSS 22: *Testy dokładne* podane są wzory umożliwiające obliczenie przybliżonej wartości statystyki chi-kwadrat na podstawie prawdopodobieństwa w teście Fishera), a w programie STATA pojawia się tylko prawdopodobieństwo. Jak widzieliśmy w podrozdziale 4.3.1.3, właśnie prawdopodobieństwo jest wynikiem testowania. Jak wykorzystać prawdopodobieństwo do obliczenia miernika wielkości efektu?

Tabela 4.19. Wyniki oceny niezależności zmiennych „grupa" i „hobby" w programie STATA

```
. tabulate hobby grupa if plec==2, chi2 exact V

Enumerating sample-space combinations:
stage 3: enumerations = 1
stage 2: enumerations = 2
stage 1: enumerations = 0

|          grupa
hobby |      1     2     3 |  Total
-----------+---------------------------------+----------
1 |        0     7     8 |     15
2 |        2     4     3 |      9
-----------+---------------------------------+----------
Total |      2    11    11 |     24

Pearson chi2(2) =  3.8303  Pr = 0.147
Cramér's V =      0.3995
Fisher's exact =                 0.250
```

Zwróćmy jeszcze uwagę na prawdopodobieństwa w obu omawianych testach. W teście chi-kwadrat niezależności prawdopodobieństwo jest równe 0,173 (wartość dokładna w teście chi-kwadrat; w wynikach z programu STATA dla chi-kwadrat podawane jest prawdopodobieństwo przybliżone), 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 jeden z testów użyty jest nieprawidłowo) 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 w przypadku tak mało licznej próby i liczebności oczekiwanych w czterech komórkach mniejszych od 5 mamy jednak prawo używać przybliżenia chi-kwadrat? A takim jest wartość 3,272.

Czy decyzje podjęte na podstawie testowania hipotez i szacowania wielkości efektu są sprzeczne? Nie można na to jednoznacznie odpowiedzieć. 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 się nie pojawiły. Próby zbudowania nowego paradygmatu idą raczej w kierunku wykorzystania pojęcia wiarygodności (Blume, 2002; Royall, 2000a; 2000b) lub ewentualnie rozwiązań bayesowskich. Do zagadnienia tego powrócę w ostatnim rozdziale.

Oceny obserwowanej mocy testu dla chi-kwadrat niezależności czy dokładnego testu Fishera nie udało mi się znaleźć w żadnym z trzech używanych wcześniej programów: SPSS 24, SYSTAT 13 i STATA 13.

4.3.2. Wielkość efektu w modelach regresji liniowej

Na wstępie chciałbym wyjaśnić, skąd biorą się pewne nieścisłości i niejednoznaczności w stosowanych oznaczeniach oraz jak je interpretować. W prezentowanych poniżej 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, że prawdopodobieństwo to jest równe 0 – ono jest mniejsze od 0,0005. Także kolumna zatytułowana „Istotność" może prowadzić do nieporozumień – w tej kolumnie znajduje się prawdopodobieństwo obliczone w teście, które jest porównywane z poziomem istotności testu.

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 mi 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 również pamiętać o uwagach Bruce'a 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 nic nie znaczy, a 0,5 dla próby stuelementowej 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. Przestrzegają 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, to powinny one być wykorzystywane z dużą ostrożnością. Ich najpoważniejszą ułomnością jest to, że 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.

Warto w tym miejscu zacytować fragment rozdziału dziewiątego książki Cohena (1988), w którym widać ogromną fascynację autora możliwościami obliczeniowymi komputerów. Fascynacja ta przenosi się także na możliwości (prawie nieograniczone) metod statystycznych.

> W czasie ostatniej dekady gwałtowny rozwój rewolucji komputerowej i wzrost zaawansowania metod statystycznych i planów badawczych w naukach behawioralnych doprowadziły do zrozumienia, że regresja wielokrotna i analiza korelacji (MRC), jako nadzwyczaj elastyczna procedura analizy danych, wyjątkowo nadaje się do rozwiązywania różnorodnych typów problemów spotykanych w badaniach behawioralnych. […] W tym „nowym spojrzeniu" […] model MRC jest bardzo ogólnym systemem analizy danych, który może być zastosowany zawsze, gdy badana jest ilościowa „zmienna zależna" (*Y*) w powiązaniu z jednym albo wieloma badanymi czynnikami, gdzie każdy z badanych czynników (*A*, *B* itp.) jest zbiorem utworzonym z jednej albo wielu „zmiennych niezależnych" […]. Forma zależności nie jest niczym ograniczona – może to być zależność prostoliniowa albo krzywoliniowa, bezwarunkowa albo warunkowa, całościowa albo cząstkowa. Nieważna jest też natura badanych czynników – mogą one być ilościowe albo jakościowe (skale nominalne), mogą być efektami głównymi albo interakcjami, zmiennymi będącymi głównym obszarem zainteresowania albo kowariancjami, zmiennymi zakłócającymi (jak w analizie kowariancji). Badane czynniki lub ich składowe zmienne niezależne mogą być skorelowane nawzajem ze sobą albo nieskorelowane (jak w planach czynnikowych dyskutowanych w poprzednim rozdziale), określające naturalnie występujące własności (cechy), takie jak płeć albo religia, albo IQ, albo – ewentualnie – poziomy poddawane manipulacjom w eksperymentach. W skrócie: praktycznie dowolna informacja może być reprezentowana jako badany czynnik i jego zależności z (albo wpływ na) *Y* badane za pomocą MRC[4] (Cohen, 1988).

4 „During the past decade, under the impetus of the computer revolution and increasing sophistication in statistics and research design among behavioral scientists, multiple regression

Lecz entuzjazm ten nie ma żadnych podstaw teoretycznych i jest całkowicie nieuzasadniony. Można go potraktować najwyżej jako pobożne życzenie.

Wymienionym wyżej wartościom współczynnika korelacji odpowiadają następujące wartości współczynnika determinacji: 0,01, 0,09 i 0,25.

W książce Cohena (1988: 410–413) proponowany jest także inny miernik wielkości efektu w wielozmiennowych modelach regresji liniowej – f^2:

$$f^2 = \frac{R^2}{1-R^2} \tag{4.16}$$

gdzie f^2 jest prostą funkcją współczynnika determinacji, przedziały dla f^2 są pochodnymi granic przedziałów dla R^2 i układają się następująco: mała wielkość efektu: $f^2 = 0{,}01$, średnia: $f^2 = 0{,}10$, duża: $f^2 = 0{,}33$.

Wzór (4.16) można łatwo przekształcić, aby uzyskać wartość R^2 na podstawie f^2:

$$R^2 = \frac{f^2}{1+f^2} \tag{4.17}$$

Przyjrzyjmy się, jakie są relacje między wartościami obu mierników:

$$f^2 = 0{,}01 \Rightarrow R^2 = 0{,}0099$$

$$f^2 = 0{,}1 \Rightarrow R^2 = 0{,}0909$$

$$f^2 = 0{,}33 \Rightarrow R^2 = 0{,}248$$

$$f^2 = 0{,}5 \Rightarrow R^2 = 0{,}333$$

Natychmiast rodzi się intuicyjne pytanie: do czego jest potrzebny miernik f^2, na czym polega jego wyższość czy większa wygoda stosowania w porównaniu z R^2? Jedyną jego przewagą nad R^2 jest brak konieczności znajomości interpretacji R^2. Ale skąd wzięły się granice (raczej wartości) dla miernika f^2?

---

and correlation analysis (MRC) has come to be understood as an exceedingly flexible data-analytic procedure remarkably suited in the variety and types of problems encountered in behavioral research. […] In this 'new look', fixed model MRC is a highly general data-analytic system that can be employed whenever a quantitative 'dependent variable' (*Y*) is to be studied in its relationship to one or more research factors of interest, where each research factor (*A*, *B*, etc.) is a set made up of one or more 'independent variables' (IVs). The form of the relationship is not constrained: it may be straight-line or curvilinear, general or conditional, whole or partial. The nature of the research factors is also not constrained: they may be quantitative or qualitative (nominal scales), main effects or interactions, variates of direct interest, or covariates to be partialled (as in the analysis of covariance). Research factors or their constituent IVs may be correlated with each other or uncorrelated (as in the factorial designs discussed in the preceding chapter), naturally occurring properties like sex or religion or IQ or, alternatively, experimentally manipulated 'treatments'. In short, virtually any information may be represented as a research factor and its relationship to (or effect on) *Y* studied by MRC".

Skoro próbujemy oceniać wielkość efektu dla wielozmiennowego modelu regresji liniowej, może warto 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 4.5.

PRZYKŁAD 4.5

Przypuszczamy, że istnieje liniowa zależność 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. Do zbudowania funkcji opisującej tę zależność wykorzystałem wielozmiennowy model regresji liniowej używając metody krokowej z prawdopodobieństwem wprowadzenia zmiennej równym 0,05 i usunięcia zmiennej 0,051.

```
regression
/missing listwise
/statistics coeff outs ci(95) r anova collin tol change zpp
/criteria=pin(.05) pout(.051)
/noorigin
/dependent subiekt
/method=stepwise soc ghq_suma wrogosc zaklopot przygneb znuzenie zyczliwo napiecie wigor.
```

Tabela 4.20. Historia wprowadzania i usuwania poszczególnych zmiennych; historia budowy końcowego modelu regresyjnego[a]

Model	Zmienne wprowadzone	Zmienne usunięte	Metoda
1	przygneb	.	krokowa (kryterium: prawdopodobieństwo F-wprowadzenia <= ,050, prawdopodobieństwo F-usunięcia >= ,051).
2	SOC	.	krokowa (kryterium: prawdopodobieństwo F-wprowadzenia <= ,050, prawdopodobieństwo F-usunięcia >= ,051).
3	napiecie	.	krokowa (kryterium: prawdopodobieństwo F-wprowadzenia <= ,050, Prawdopodobieństwo F-usunięcia >= ,051).

Model	Zmienne wprowadzone	Zmienne usunięte	Metoda
4	zaklopot	.	krokowa (kryterium: prawdopodobieństwo F-wprowadzenia <= ,050, prawdopodobieństwo F-usunięcia >= ,051).
5	.	przygneb	krokowa (kryterium: prawdopodobieństwo F-wprowadzenia <= ,050, prawdopodobieństwo F-usunięcia >= ,051).
6	wrogosc	.	krokowa (kryterium: prawdopodobieństwo F-wprowadzenia <= ,050, prawdopodobieństwo F-usunięcia >= ,051).
7	GHQ_suma	.	krokowa (kryterium: prawdopodobieństwo F-wprowadzenia <= ,050, prawdopodobieństwo F-usunięcia >= ,051).
8	.	napiecie	krokowa (kryterium: prawdopodobieństwo F-wprowadzenia <= ,050, prawdopodobieństwo F-usunięcia >= ,051).

[a] Zmienna zależna: subiekt.

Tabela 4.21. Podsumowanie kolejnych kroków modelowania zależności liniowej

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
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, wrogosc.
[g] Predyktory: (Stała), SOC, napiecie, zaklopot, wrogosc, GHQ_suma.
[h] Predyktory: (Stała), SOC, zaklopot, wrogosc, GHQ_suma.

Tabela 4.22. Wyniki testowania hipotezy określonej wzorem (4.18) w kolejnych krokach budowy modelu[a]

Model		Suma kwadratów	*df*	Średni kwadrat	*F*	Istotność
1	regresja	139 349,556	1	139 349,556	233,653	,000[b]
	reszta	261 220,987	438	596,395		
	ogółem	400 570,543	439			
2	regresja	163 156,368	2	81 578,184	150,158	,000[c]
	reszta	237 414,175	437	543,282		
	ogółem	400 570,543	439			
3	regresja	171 568,514	3	57 189,505	108,884	,000[d]
	reszta	229 002,029	436	525,234		
	ogółem	400 570,543	439			
4	regresja	174 812,271	4	43 703,068	84,209	,000[e]
	reszta	225 758,272	435	518,985		
	ogółem	400 570,543	439			
5	regresja	173 572,124	3	57 857,375	111,128	,000[f]
	reszta	226 998,419	436	520,639		
	ogółem	400 570,543	439			
6	regresja	177 598,699	4	44 399,675	86,620	,000[g]
	reszta	222 971,844	435	512,579		
	ogółem	400 570,543	439			
7	regresja	180 706,944	5	36 141,389	71,341	,000[h]
	reszta	219 863,599	434	506,598		
	ogółem	400 570,543	439			
8	regresja	179 823,093	4	44 955,773	88,589	,000[i]
	reszta	220 747,450	435	507,465		
	ogółem	400 570,543	439			

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

$$\begin{cases} H_0: R^2 = 0 \\ H_1: R^2 > 0 \end{cases} \tag{4.18}$$

Prawdopodobieństwo w każdym modelu jest mniejsze od 0,0005. Końcowy model powstał po ośmiu krokach:

Model 1: do modelu zostaje wprowadzona zmienna „przygneb" (przygnębienie).

Model 2: do modelu włączona zostaje zmienna „SOC" (poczucie koherencji).

Model 3: do zmiennych „przygnębienie" i „SOC" dołączona zostaje zmienna „napięcie".

Model 4: w czwartym kroku włączona zostaje zmienna „zaklopot" (zakłopotanie).

Model 5: włączenie do modelu zmiennej „zakłopotanie" powoduje „utratę istotności" przez zmienną „przygneb" (przygnębienie), która w piątym kroku zostaje usunięta z modelu; powstaje model trzyzmiennowy: „SOC", „napięcie" i „zaklopot" (zakłopotanie).

Model 6: do trzyzmiennowego modelu 5 dołączona zostaje zmienna „wrogosc" (wrogość).

Model 7: do modelu włączona zostaje zmienna „GHQ_suma" (subiektywna ocena stanu zdrowia).

Model 8: włączenie do modelu zmiennej „GHQ_suma" powoduje „utratę istotności" przez zmienną „napięcie".

Ostateczny model (model ze zmienną objaśniającą „subiekt" i zaproponowanym wyjściowym zespołem zmiennych objaśniających) jest modelem z czterema zmiennymi objaśniającymi. Zmienne: „znuzenie", „życzliwo" (życzliwość), „napiecie" i „przygneb" (przygnębienie) zostały poza modelem jako zmienne „nieistotne statystycznie".

Przyjrzyjmy się teraz standaryzowanym współczynnikom regresji (β – beta). W tab. 4.23 pokazuję tylko ostatni, końcowy model.

Tabela 4.23. Współczynniki regresji i ich ocena w końcowym modelu[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.

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, a dla zmiennej GHQ_suma 0,148. Niby jest to sensowne, gdyż każda następna zmienna wprowadzana do modelu mniej do niego wnosi (mierzone R^2), 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; beta zmiennej wrogość jest nieznacznie większa: 0,193. Ale czy możemy tutaj mówić o jakiejkolwiek ocenie wielkości efektu? Podobne zastrzeżenia do wykorzystywania standaryzowanych współczynników regresji jako miar wielkości efektu mają Greenland i wsp. (1986, 1991). Mimo że ich obiekcje dotyczą tego typu mierników wielkości efektów w zagadnieniach biologicznych i 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 (moja „mantra").

Warto zwrócić uwagę na relacje między R^2 i F:

$$R^2 = \frac{SS_{reg}}{SS_Y} = \frac{SS_{reg}}{SS_{reg}+SS_{res}} \tag{4.19}$$

$$F = \frac{MS_{reg}}{MS_{res}} = \frac{SS_{reg}/df_{reg}}{SS_{res}/df_{res}} \tag{4.20}$$

Wartość współczynnika determinacji R^2 nie uwzględnia w bezpośredni sposób liczb stopni swobody dla odpowiednich sum kwadratów, co ma miejsce przy obliczaniu wartości statystki *F*-Snedecora. Mimo pozornego podobieństwa wzorów są to jednak różne mierniki.

Wartość współczynnika determinacji z próby dla modelu 8 (modelu bez „nieistotnych" zmiennych objaśniających) wynosi 0,449, a wartość miernika f^2 jest równa 0,815. Czy wartość f^2 więcej wnosi do naszej wiedzy o merytorycznej przydatności, ważności, znaczeniu modelu niż wartość R^2? To, że jest większa od wartości R^2 o niczym nie świadczy; $f^2 > 0{,}33$, czyli według Cohena jest to duża wielkość efektu. Zamieniamy po prostu jeden miernik na inny.

4.3.3. Wielkość efektu w modelach analizy wariancji

W podręczniku Tabachnick i Fidell (2007), a także w wielu artykułach (np. Bakeman, 2005; Levine, Hullett, 2002; Richardson, 2011) znajdujemy mierniki wielkości efektu wykorzystywane w modelach analizy wariancji, zarówno w modelu jednoczynnikowej jednozmiennowej analizy wariancji, jak i w modelach z powtarzanymi obserwacjami. Podstawowe są: współczynnik η^2 wyrażony wzorem:

$$\eta^2 = \frac{SS_{effect}}{SS_{total}} \tag{4.21}$$

oraz cząstkowy współczynnik η_p^2:

$$\eta_p^2 = \frac{SS_{effect}}{SS_{effect}+SS_{error}} \tag{4.22}$$

Rzadziej używany jest współczynnik $\hat{\omega}^2$:

$$\hat{\omega}^2 = \frac{SS_{effect}-(df_{effect})\cdot MS_{error}}{SS_{total}+MS_{error}} \tag{4.23}$$

gdzie:

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

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

SS_{error} – to zmienność spowodowana błędem eksperymentalnym, jest 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} – średnia SS_{error}; $MS_{error} = SS_{error}/df_{error}$,
df – liczba 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). Występowanie w planie badawczym większej liczby efektów minimalizuje miernik cząstkowej η^2. **Uwaga**: mierniki η^2 i η_p^2 w jednoczynnikowej analizie wariancji są jednakowe. W innych modelach analizy wariancji $\eta^2 < \eta_p^2$, co wynika z porównania wzorów (4.22) i (4.23).

Wróćmy na moment do programu SPSS. Program ten 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 również liczone są cząstkowe η^2.

Oczywiście, można obliczyć wartości mierników, 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 η_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 te zaproponowane przez Cohena. Może przyjęcie granic Cohena dla η^2, zaś propozycji Sinka i Mvududu jako granic dla η_p^2 byłoby sensownym rozwiązaniem, ale nigdzie nie znalazłem takiej propozycji.

Zaproponowane powyżej granice w postaci, w jakiej są przedstawione, praktycznie uniemożliwiają ich wykorzystanie. Efekt słaby to $\eta^2 = 0{,}01$, a jak zinterpretować wartość $\eta^2 = 0{,}011$? Proponowanie izolowanych wartości miernika η^2 jest całkowicie pozbawione sensu, gdyż konieczne jest podanie przedziałów, które będą charakteryzowały siłę zależności. Proponowałbym zinterpretować to następująco:

Tabela 4.24. Przedziały dla miernika η^2 i ich interpretacja werbalna

Werbalna ocena wielkości efektu	Wartość miernika η^2	
	według Cohena (1988)	według Sinka i Mvududu (2010)
Brak efektu	$\eta^2 < 0{,}01$	$\eta^2 < 0{,}01$
Efekt słaby	$0{,}01 \leq \eta^2 < 0{,}09$	$0{,}01 \leq \eta^2 < 0{,}06$
Efekt umiarkowany	$0{,}09 \leq \eta^2 < 0{,}25$	$0{,}06 \leq \eta^2 < 0{,}14$
Efekt silny	$\eta^2 \geq 0{,}25$	$\eta^2 \geq 0{,}14$

Pewien niepokój budzi duży przedział wartości η^2 oznaczający silny efekt (teoretyczny przedział zmienności η^2; w tym przypadku to [0,25; 1] w interpretacji Cohena oraz [0,14; 1] w interpretacji Sinka i Mvududu).

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

Poniżej kilka przykładów, w których wykorzystano program SPSS.

PRZYKŁAD 4.6

Porównujemy wartości oczekiwane zmiennej „subiekt" (subiektywne odczucie stresu związanego z pracą zawodową) w grupach zawodowych: strażaków, pracowników służb więziennych i policjantów. Kategorie zmiennej grupa: 1 → strażacy, 2 → pracownicy służb więziennych, 3 → policjanci.

```
ONEWAY subiekt BY grupa
/STATISTICS DESCRIPTIVES HOMOGENEITY BROWNFORSYTHE WELCH
/MISSING ANALYSIS
/POSTHOC=BONFERRONI T3 ALPHA(0.05).
```

Jednoczynnikowa analiza wariancji (ONEWAY)

Tabela 4.25. Statystyki opisowe zmiennej „subiekt" w grupach zawodowych

	N	Średnia	Odchylenie standardowe	Błąd standardowy	95% przedział ufności dla średniej		Minimum	Maksimum
					dolna granica	górna granica		
1,0	101	108,198	23,6787	2,3561	103,524	112,872	68,0	211,0
2,0	90	118,622	33,5414	3,5356	111,597	125,647	73,0	231,0
3,0	253	115,494	30,9438	1,9454	111,663	119,325	63,0	229,0
Ogółem	444	114,468	30,1718	1,4319	111,654	117,283	63,0	231,0

Tabela 4.26. Wyniki testu Levene'a oceny równości wariancji zmiennej „subiekt"

Test Levene'a	*df*1	*df*2	Istotność
7,180	2	441	,001

Tabela 4.27. Wyniki porównania wartości oczekiwanych zmiennej „subiekt" testem *F*-Snedecora

	Suma kwadratów	*df*	Średni kwadrat	*F*	Istotność
Między grupami	5 790,122	2	2 895,061	3,212	,041
Wewnątrz grup	397 490,436	441	901,339		
Ogółem	403 280,559	443			

Tabela 4.28. Wyniki porównania wartości oczekiwanych zmiennej „subiekt" odpornymi testami Welcha i Browna-Forsythe'a

	Statystyka[a]	*df*1	*df*2	Istotność
Welch	4,093	2	199,673	,018
Brown-Forsythe	3,324	2	261,845	,038

[a] Rozkład *F* asymptotyczny.

W tab. 4.27 przedstawione zostały sumy kwadratów:

$$SS_{effect} = SS_{między\ grupami} = 5790{,}122$$

$$SS_{total} = SS_{ogółem} = 403280{,}559$$

$$\eta^2 = \eta_p^2 = \frac{5790{,}112}{403280{,}559} = 0{,}014 \tag{4.25}$$

zatem zgodnie z propozycjami Cohena (1988) uznajemy efekt za słaby. Można by się z tym zgodzić, gdyż pewien efekt między zmienną „grupa" i zmienną „subiekt" istnieje; prawdopodobieństwo w teście *F*-Snedecora jest równe 0,041, a więc jest mniejsze od 0,05, choć bliskie przyjmowanego zazwyczaj poziomu istotności 0,05. W tym momencie z pełną świadomością korzystam z mieszaniny teorii testowania hipotez Fishera i Neymana-Pearsona, być może z jeszcze innymi „zanieczyszczeniami", i porównuję prawdopodobieństwa. Dla przypomnienia, w tab. 4.26 przedstawione są wyniki testowania hipotez:

$$\begin{cases} H_0: \mu_1 = \mu_2 = \mu_3 \\ H_1: \neg(\mu_1 = \mu_2 = \mu_3) \end{cases} \tag{4.26}$$

Na podstawie wyników testu *F*-Snedecora odrzucamy hipotezę zerową i przyjmujemy alternatywną (ostatnie zdanie jest w pełni zgodne z teorią testowania hipotez statystycznych Neymana-Pearsona).

Jednak istotny problem pojawia się w innym miejscu. Aby móc skorzystać z wyników przedstawionych w tab. 4.27, czyli z klasycznego testu *F*-Snedecora, musi być spełnione założenie o jednorodności wariancji zmiennej „subiekt" w porównywanych grupach. Założenie to nie jest spełnione, gdyż prawdopodobieństwo w teście Levene'a jest równe 0,001. W tym momencie powinniśmy skorzystać z odpornych testów Welcha oraz Browna-Forsythe'a. Ale w tej tabelce nie mamy sum kwadratów. Czy są one takie same jak w tab. 4.27? Nie jestem pewien, gdyż testy Welcha i Browna-Forsythe'a wykorzystują asymptotyczny rozkład *F*-Snedecora. Może więc nie warto, używając oceny wielkości efektu, niczego zakładać o badanych próbkach, tylko liczyć wielkość efektu? Problem w tym, że zasady testowania hipotez w teorii Neymana-Pearsona mają dość solidne podstawy teoretyczne, natomiast ocena wielkości efektu to pewna proteza, mająca zastąpić konieczność wyjaśnienia przez badacza uzyskanych wyników w terminach merytorycznych, gdy rozwiązanie statystyczne jest niewystarczające albo niejednoznaczne. W rzeczywistości jednak ocena wielkości efektu w swej najprostszej postaci wykorzystuje związek między wartością statystyki *t*-Studenta i wartością współczynnika korelacji z próby, czyli wykorzystuje twierdzenie stosowane przy testowaniu hipotez w teorii Neymana-Pearsona. Zwalniamy się z myślenia, gdyż z obliczeń wyszło, że efekt jest słaby! BRAWO!

Relacje między η^2 i F w modelu jednoczynnikowej jednozmiennowej analizy wariancji są takie same, jak relacje między R^2 i F w modelach regresji liniowej (wzory (4.19) i (4.20)):

$$\eta^2 = \frac{SS_{\text{między grupami}}}{SS_{\text{ogółem}}} = \frac{SS_{\text{między grupami}}}{SS_{\text{między grupami}} + SS_{\text{wewnątrz grup}}} \tag{4.27}$$

$$F = \frac{MS_{\text{między grupami}}}{MS_{\text{wewnątrz grup}}} = \frac{SS_{\text{między grupami}}/df_{\text{między grupami}}}{SS_{\text{wewnątrz grup}}/df_{\text{wewnątrz grup}}} \tag{4.28}$$

PRZYKŁAD 4.7

Włączmy do modelu opisującego zależność między zmienną „grupa" (grupa zawodowa) i zmienną „subiekt" (subiektywną oceną stresu związanego z pracą) zmienną kowariancyjną „SOC" (poczucie koherencji). Nie można wykluczyć, że na subiektywne odczucie stresu związanego z pracą ma wpływ nie tyle grupa zawodowa, ile poczucie koherencji pracowników odpowiedniej profesji.

Sprawdźmy najpierw, czy pomysł potraktowania zmiennej „SOC" jako zmiennej kowariancyjnej w modelu jednoczynnikowej jednozmiennowej analizy wariancji ma sens od strony statystycznej.

Tabela 4.29. Wyniki testu Levene'a oceny równości wariancji zmiennej „SOC"

Test Levene'a	*df*1	*df*2	Istotność
3,329	2	441	,037

Tabela 4.30. Wyniki porównania wartości oczekiwanych zmiennej „SOC" odpornymi testami Welcha i Browna-Forsythe'a

Test	Statystyka[a]	*df*1	*df*2	Istotność
Welch	4,618	2	182,799	,011
Brown-Forsythe	3,945	2	247,649	,021

[a] Rozkład *F* asymptotyczny.

Ponieważ wariancje zmiennej „SOC" nie są jednorodne w grupach zawodowych, do porównywania wartości oczekiwanych zostały wykorzystane testy Welcha i Browna-Forsythe'a. Prawdopodobieństwa w obu testach są mniejsze od 0,05, zatem uznajemy, że prawdziwe średnie nie są jednakowe w grupach zawodowych.

Tabela 4.31. Współczynniki regresji i ich ocena w modelu opisującym zależność liniową między zmiennymi „subiekt" i „SOC"

Model		Współczynniki niestandaryzowane		Współczynniki standaryzowane	*t*	Istotność	95,0% przedział ufności dla B	
		B	błąd standardowy	Beta			dolna granica	górna granica
1	(stała)	223,546	7,360		30,372	,000	209,080	238,011
	SOC	–,752	,050	–,581	–15,010	,000	–,851	–,654

Istnieje liniowa zależność między zmiennymi „SOC" i „subiekt". Wyniki zawarte w tab. 4.30 i 4.31 pozwalają uznać zmienną „SOC" za zmienną kowariancyjną w modelu analizy wariancji. Poniżej przedstawiona jest realizacja analizy kowariancji.

```
UNIANOVA subiekt BY grupa WITH SOC
/METHOD=SSTYPE(3)
/INTERCEPT=INCLUDE
/EMMEANS=TABLES(grupa) WITH(SOC=MEAN) COMPARE ADJ(BONFERRONI)
/PRINT=ETASQ DESCRIPTIVE HOMOGENEITY OPOWER
/CRITERIA=ALPHA(.05)
/DESIGN=SOC grupa.
```

Tabela 4.32. Wyniki testów efektów międzyobiektowych dla zmiennej „subiekt" ze zmienną kowariancyjną „SOC"

Źródło	Typ III sumy kwadratów	*df*	Średni kwadrat	*F*	Istotność	Cząstkowe eta kwadrat	Parametr niecentralności	Moc obserwowana[b]
Model skorygowany	136980,650[a]	3	45660,216	75,443	,000	,340	226,329	1,000
Stała	535137,754	1	535137,754	884,193	,000	,668	884,193	1,000
SOC	131190,527	1	131190,527	216,762	,000	,330	216,762	1,000
Grupa	826,517	2	413,259	,683	,506	,003	1,366	,165
Błąd	266299,909	440	605,227					
Ogółem	6221026,000	444						
Ogółem skorygowane	403280,559	443						

[a] *R* kwadrat = ,340 (skorygowane *R* kwadrat = ,335).
[b] Obliczone z użyciem alfa = ,05.

Wróćmy na moment do podstaw teoretycznych jednoczynnikowej analizy kowariancji. Postać modelu jest, w rozważanej przez nas sytuacji, następująca:

$$y_{ij} = \mu + \alpha_i + \beta X_{ij} + \varepsilon_{ij} \tag{4.29}$$

gdzie:
μ – ogólna średnia badanej populacji,
α_i – efekt i-tej kategorii zmiennej dyskretnej (zabieg, *treatment*),
ε_{ij} – błąd eksperymentalny,
X_{ij} – wartości zmiennej kowariancyjnej,
β – współczynnik regresji.

Testowane są hipotezy:

$$\begin{cases} H_0: \alpha_1 = \alpha_2 = \cdots = \alpha_k \\ H_1: \neg(\alpha_1 = \alpha_2 = \cdots = \alpha_k) \end{cases} \text{ albo } \begin{cases} H_0: \mu_1 = \mu_2 = \cdots = \mu_k \\ H_1: \neg(\mu_1 = \mu_2 = \cdots = \mu_k) \end{cases} \tag{4.30}$$

oraz

$$\begin{cases} H_0: \beta = 0 \\ H_1: \beta \neq 0 \end{cases} \tag{4.31}$$

Prawdopodobieństwo w teście dotyczącym współczynnika regresji (SOC w tab. 4.32) jest mniejsze od 0,0005, a $\eta_p^2 = 0{,}330$, zaś dla hipotezy o równości wartości oczekiwanych zmiennej „subiekt" w grupach zawodowych $p = 0{,}506$, a $\eta_p^2 = 0{,}003$ i obserwowana moc testu jest równa 0,165. Oznacza to, iż prawdopodobieństwo błędu drugiego rodzaju wynosi 0,835. Na podstawie prawdopodobieństw uzyskanych w testach (kolumna „Istotność" w tab. 4.32) wiemy, że „SOC" jest istotną zmienną kowariancyjną, tj. w istotny sposób modyfikuje relację między zmiennymi „grupa" i „subiekt", natomiast $p = 0{,}506$ dla zmiennej „grupa" oznacza, iż nie możemy odrzucić hipotezy zerowej o równości wartości oczekiwanych zmiennej „subiekt" w grupach zawodowych. Wykorzystując $\eta_p^2 = 0{,}003$ stwierdzamy, że nie istnieje zależność między zmiennymi „grupa" i „subiekt", ale obserwowana moc testu jest tak mała, że hipotezy zerowej nie możemy przyjąć. O czym zatem „mówi" η_p^2? Czy używając miernika η_p^2 dowiedzieliśmy się czegoś więcej, niż na podstawie prawdopodobieństw w testach? I pytanie istotniejsze: czy wartości η_p^2 „mówią" o tym samym w przypadku zmiennej „SOC" i zmiennej „grupa"? Czy będziemy je mogli kiedykolwiek porównywać?

W dwóch kolejnych przykładach, 4.8 i 4.9, pojawia się pojęcie średnich brzegowych. Jest to terminologia stosowana w SPSS, a oznacza średnie skorygowane o wpływ zmiennej kowariancyjnej, w tym przypadku o SOC. Oto przykład obrazujący dokładniej problem porównywania różnych wartości η_p^2.

PRZYKŁAD 4.8

Oceńmy zależność między zmienną „grupa_wieku", gdzie grupy wieku są określone jako:

$$\text{grupa\_wieku} = \begin{cases} 1 & \text{wiek} \leq 35 \text{ lat} \\ 2 & 35 < \text{wiek} \leq 45 \text{ lat} \\ 3 & \text{wiek} > 45 \text{ lat} \end{cases}$$

i zmienną „GHQ_suma" (subiektywna ocena stanu zdrowia na podstawie wyników kwestionariusza Goldberga zawierającego 28 pytań) w obecności zmiennej kowariancyjnej „SOC". Poniżej przedstawiam fragmenty wydruku z programu SPSS.

```
UNIANOVA GHQ_suma BY grupa_wieku WITH SOC
/METHOD=SSTYPE(3)
/INTERCEPT=INCLUDE
/EMMEANS=TABLES(grupa_wieku) WITH(SOC=MEAN) COMPARE ADJ(BONFERRONI)
/PRINT=ETASQ HOMOGENEITY DESCRIPTIVE OPOWER
/CRITERIA=ALPHA(.05)
/DESIGN=SOC grupa_wieku.
```

Tabela 4.33. Statystyki opisowe (średnia, odchylenie standardowe i liczebność grupy) zmiennej „GHQ_suma" w grupach wiekowych (są to wyniki surowe, tzn. bez uwzględnienia zmiennej kowariancyjnej)

grupa_wieku	Średnia	Odchylenie standardowe	*N*
1	17,054	7,2372	185
2	22,932	10,7988	190
3	26,590	12,8923	61
Ogółem	20,950	10,4105	436

Tabela 4.34. Wyniki testu Levene'a oceny równości wariancji zmiennej „GHQ_suma"[a]

F	*df*1	*df*2	Istotność
4,847	2	433	,008

Testuje hipotezę zerową zakładającą, że wariancja błędu zmiennej zależnej jest równa we wszystkich grupach.

[a] Plan: stała + SOC + grupa_wieku.

Tabela 4.35. Wyniki testów efektów międzyobiektowych dla zmiennej „GHQ_suma" w grupach wieku, ze zmienną kowariancyjną „SOC"

Źródło	Typ III sumy kwadratów	*df*	Średni kwadrat	*F*	Istotność	Cząstkowe eta kwadrat	Parametr niecentralności	Moc obserwowana[b]
Model skorygowany	22 853,574[a]	3	7 617,858	135,477	,000	,485	406,431	1,000
Stała	41 086,027	1	41 086,027	730,679	,000	,628	730,679	1,000
SOC	17 359,008	1	17 359,008	308,715	,000	,417	308,715	1,000
grupa_wieku	3 162,533	2	1 581,267	28,121	,000	,115	56,243	1,000
Błąd	24 291,316	432	56,230					
Ogółem	238 498,000	436						
Ogółem skorygowane	47 144,890	435						

[a] *R* kwadrat = ,485 (skorygowane *R* kwadrat = ,481).

[b] Obliczone z użyciem alfa = ,05.

Tabela 4.36. Statystyki opisowe zmiennej „GHQ_suma" skorygowanej o wpływ zmiennej „SOC"[5] w grupach wiekowych

grupa_wieku	Średnia	Błąd standardowy	95% przedział ufności	
			dolna granica	górna granica
1	18,088[a]	,554	16,998	19,178
2	22,194[a]	,546	21,122	23,267
3	25,751[a]	,961	23,862	27,640

[a] Współzmienne występujące w modelu zostały oszacowane jako następujące wartości: SOC = 144,679.

Tabela 4.37. Wyniki porównań parami wartości oczekiwanych zmiennej „GHQ_suma" skorygowanej o wpływ zmiennej kowariancyjnej „SOC" między grupami wiekowymi

(I) grupa_wieku	(J) grupa_wieku	Różnica średnich (I–J)	Błąd standardowy	Istotność[b]	95% przedział ufności dla różnicy[b]	
					dolna granica	górna granica
1	2	–4,106*	,781	,000	–5,983	–2,229
	3	–7,663*	1,112	,000	–10,336	–4,990
2	1	4,106*	,781	,000	2,229	5,983
	3	–3,557*	1,104	,004	–6,209	–,905
3	1	7,663*	1,112	,000	4,990	10,336
	2	3,557*	1,104	,004	,905	6,209

Na podstawie estymowanych średnich brzegowych.
* Różnica średnich jest istotna na poziomie ,05.
[b] Poprawka dla porównań wielokrotnych – Bonferroniego.

Przy ocenie „istotności" zmiennych „SOC" i „grupa_wieku" w tabeli prezentującej wyniki testów efektów międzyobiektowych (tab. 4.35) obliczane są wartości mierników: cząstkowe eta kwadrat i obserwowana moc testu. Natomiast dla testów porównań wielokrotnych (tab. 4.37) „porównania parami" wartości tych mierników już nie są obliczane. Problemy w „teorii" czy oprogramowaniu?

Pojawia się jeszcze jeden problem: czy możemy użyć jednozmiennowej analizy kowariancji, skoro prawdopodobieństwo w teście Levene'a jest równe 0,008, co oznacza, iż wariancje zmiennej „GHQ_suma" nie są równe we wszystkich grupach

5 Średnie analizowanej zmiennej skorygowanej o wpływ zmiennej kowariancyjnej, w programie SPSS nazywane są średnimi brzegowymi.

wiekowych? Możemy jednak skorzystać z konwencji wykorzystywanej przy braku jednorodności wariancji badanej zmiennej w grupach (podrozdział 2.3). Krotność między największym i najmniejszym odchyleniem standardowym z próby jest równa 12.892/7,237 = 1,78, czyli jest mniejsza od 2. Możemy więc uznać, że różnice między odchyleniami standardowymi są na tyle nieduże, iż wyniki analizy wariancji będą poprawne (odporność statystyki *F*-Snedecora na niespełnianie założenia jednorodności wariancji).

Tabela 4.38. Porównanie wartości surowych parametrów rozkładu prawdopodobieństwa badanej cechy i wartości parametrów skorygowanych o wpływ zmiennej kowariancyjnej „SOC"

grupa_ wieku	*n*	Parametry surowe		Parametry zmiennej skorygowanej		
		M	SD	M_adj	SE_adj	SD_adj
1	185	17,054	7,237	18,088	0,554	7,535
2	190	22,932	10,799	22,194	0,546	7,526
3	61	26,590	12,892	25,751	0,961	7,506

PRZYKŁAD 4.9

Do modelu jednoczynnikowej analizy kowariancji z przykładu 4.8 dodano drugi czynnik, zmienną dwustanową „ukl_kraz". Wartość zmiennej ukl_kraz = 1 oznacza grupę osób, u których zdiagnozowano jakąś chorobę układu krążenia, zaś ukl_kraz = 0 to grupa osób, u których nie stwierdzono chorób układu krążenia. Uzyskano w ten sposób dwuczynnikowy jednozmiennowy model analizy kowariancji ze zmienną „SOC" jako kowariancją.

```
UNIANOVA GHQ_suma BY grupa_wieku ukl_kraz WITH SOC
/METHOD=SSTYPE(3)
/INTERCEPT=INCLUDE
/EMMEANS=TABLES(grupa_wieku) WITH(SOC=MEAN) COMPARE ADJ(BONFERRONI)
/EMMEANS=TABLES(ukl_kraz) WITH(SOC=MEAN) COMPARE ADJ(BONFERRONI)
/EMMEANS=TABLES(grupa_wieku*ukl_kraz) WITH(SOC=MEAN) COMPARE(grupa_wieku)
ADJ(BONFERRONI)
/EMMEANS=TABLES(grupa_wieku*ukl_kraz) WITH(SOC=MEAN) COMPARE(ukl_kraz)
ADJ(BONFERRONI)
/PRINT=ETASQ DESCRIPTIVE HOMOGENEITY OPOWER
/CRITERIA=ALPHA(.05)
/DESIGN=SOC grupa_wieku ukl_kraz grupa_wieku*ukl_kraz.
```

Tabela 4.39. Statystyki opisowe zmiennej „GHQ_suma" w grupach wyznaczonych przez zmienne grupa_wieku × ukl_kraz (dane surowe, bez uwzględnienia kowariancji)

grupa_wieku	ukl_kraz	Średnia	Odchylenie standardowe	*N*
1	,0	16,411	6,6225	151
	1,0	19,912	9,0767	34
	ogółem	17,054	7,2372	185
2	,0	21,691	10,0728	123
	1,0	25,209	11,7583	67
	ogółem	22,932	10,7988	190
3	,0	22,724	12,9032	29
	1,0	30,094	12,0278	32
	ogółem	26,590	12,8923	61
Ogółem	,0	19,158	9,2637	303
	1,0	25,030	11,6911	133
	ogółem	20,950	10,4105	436

Tabela 4.40. Wyniki testu Levene'a oceny równości wariancji zmiennej „GHQ_suma" w grupach wieku[a]

F	*df*1	*df*2	Istotność
3,666	5	430	,003

Testuje hipotezę zerową zakładającą, że wariancja błędu zmiennej zależnej jest równa we wszystkich grupach.

[a] Plan: stała + SOC + grupa_wieku + ukl_kraz + grupa_wieku * ukl_kraz.

Tabela 4.41. Wyniki testów efektów międzyobiektowych dla zmiennej „GHQ_suma" w grupach wyznaczonych przez zmienne grupa_wieku × ukl_kraz, ze zmienną kowariancyjną „SOC"

Źródło	Typ III sumy kwadratów	*df*	Średni kwadrat	*F*	Istotność	Cząstkowe eta kwadrat	Parametr niecentralności	Moc obserwowana[b]
Model skorygowany	23505,888[a]	6	3917,648	71,097	,000	,499	426,584	1,000
Stała	39828,627	1	39828,627	722,809	,000	,628	722,809	1,000
SOC	16308,123	1	16308,123	295,959	,000	,408	295,959	1,000

Źródło	Typ III sumy kwadratów	*df*	Średni kwadrat	*F*	Istotność	Cząstkowe eta kwadrat	Parametr niecentralności	Moc obserwowana[b]
grupa_wieku	2101,093	2	1050,546	19,065	,000	,082	38,131	1,000
ukl_kraz	586,672	1	586,672	10,647	,001	,024	10,647	,902
grupa_wieku * ukl_kraz	45,344	2	22,672	,411	,663	,002	,823	,117
Błąd	23639,002	429	55,103					
Ogółem	238498,000	436						
Ogółem skorygowane	47144,890	435						

[a] *R* kwadrat = ,499 (skorygowane *R* kwadrat = ,492).
[b] Obliczone z użyciem alfa = ,05.

Prawdopodobieństwo w teście hipotez:

$$\begin{cases} H_0: \beta = 0 \\ H_1: \beta \neq 0 \end{cases} \tag{4.32}$$

dla zmiennej „SOC" jest mniejsze od 0,0005, a $\eta_p^2 = 0{,}408$, co oznacza, że efekt jest silny. Ale efekt między czym a czym? Wszak zmienna „SOC" jest zmienną kowariancyjną, zmienną zakłócającą relacje między czynnikami a badaną zmienną ciągłą. Dla zmiennej „grupa_wieku" prawdopodobieństwo w teście hipotezy (dla tej zmiennej użyłem w indeksie oznaczenia α):

$$\begin{cases} H_0: \mu_{1\alpha} = \mu_{2\alpha} = \mu_{3\alpha} \\ H_1: \neg(\mu_{1\alpha} = \mu_{2\alpha} = \mu_{3\alpha}) \end{cases} \tag{4.33}$$

również jest mniejsze od 0,0005 i $\eta_p^2 = 0{,}082$. Według Cohena efekt tej zmiennej jest słaby, zaś według Sinka i Mvududu (tab. 4.24) umiarkowany. Wniosek: efekt jest, lecz to już wiemy na podstawie prawdopodobieństwa w teście. Dla zmiennej „ukl_kraz" prawdopodobieństwo w teście hipotez (dla zmiennej „ukl_kraz" użyłem w indeksie oznaczenia β):

$$\begin{cases} H_0: \mu_{1\beta} = \mu_{2\beta} \\ H_1: \mu_{1\beta} \neq \mu_{2\beta} \end{cases} \tag{4.34}$$

p = 0,001, a $\eta_p^2 = 0{,}024$. Według obu interpretacji, zarówno Cohena, jak i Sinka i Mvududu (tab. 4.24), zależność jest słaba. A może mamy prawo powiedzieć, że

efekt zmiennej „ukl_kraz" jest słabszy niż zmiennej „grupa_wieku"? Chociaż należałoby się zastanowić, jakie mamy do tego przesłanki.

Prawdopodobieństwo dla oceny interakcji jest równe 0,663, $\eta_p^2 = 0{,}002$, a obserwowana moc testu 0,117. Na podstawie wartości parametru „cząstkowe eta kwadrat" uznajemy, że brak takiego efektu, jednak obserwowana moc testu nie upoważnia nas do przyjęcia hipotezy zerowej.

Tabela 4.42. Statystyki opisowe zmiennej „GHQ_suma" skorygowanej o wpływ zmiennej „SOC"[6] w grupach wyznaczonych przez zmienne: grupa_wieku × ukl_kraz (średnie brzegowe)

grupa_wieku	ukl_kraz	Średnia	Błąd standardowy	95% przedział ufności	
				dolna granica	górna granica
1	,0	17,716[a]	,609	16,519	18,913
	1,0	19,613[a]	1,273	17,110	22,115
2	,0	21,265[a]	,670	19,948	22,581
	1,0	23,948[a]	,910	22,160	25,736
3	,0	23,647[a]	1,379	20,936	26,359
	1,0	27,694[a]	1,320	25,100	30,288

[a] Współzmienne występujące w modelu zostały oszacowane jako następujące wartości: SOC = 144,679.

Tabela 4.43. Wyniki porównań parami wartości oczekiwanych zmiennej „GHQ_suma" skorygowanej o wpływ zmiennej kowariancyjnej „SOC" między grupami wieku w kategoriach zmiennej „ukl_kraz"

ukl_kraz	(I) grupa_wieku	(J) grupa_wieku	Różnica średnich (I–J)	Błąd standardowy	Istotność[b]	95% przedział ufności dla różnicy[b]	
						dolna granica	górna granica
,0	1	2	–3,549*	,907	,000	–5,729	–1,368
		3	–5,931*	1,505	,000	–9,549	–2,314
	2	1	3,549*	,907	,000	1,368	5,729
		3	–2,383	1,534	,364	–6,070	1,305
	3	1	5,931*	1,505	,000	2,314	9,549
		2	2,383	1,534	,364	–1,305	6,070

6 Średnie analizowanej zmiennej, skorygowanej o wpływ zmiennej kowariancyjnej, w programie SPSS nazywane są średnimi brzegowymi.

ukl_kraz	(I) grupa_wieku	(J) grupa_wieku	Różnica średnich (I–J)	Błąd standardowy	Istotność[b]	95% przedział ufności dla różnicy[b]	
						dolna granica	górna granica
1,0	1	2	–4,335*	1,564	,017	–8,094	–,576
		3	–8,081*	1,832	,000	–12,485	–3,677
	2	1	4,335*	1,564	,017	,576	8,094
		3	–3,746	1,596	,058	–7,583	,091
	3	1	8,081*	1,832	,000	3,677	12,485
		2	3,746	1,596	,058	–,091	7,583

Na podstawie estymowanych średnich brzegowych.
* Różnica średnich jest istotna na poziomie ,05.
[b] Poprawka dla porównań wielokrotnych – Bonferroniego.

W tab. 4.44 przedstawione są wyniki testowania hipotez opisanych wzorem (4.35) w warstwach wyznaczonych przez wartości zmiennej „ukl_kraz".

$$\begin{cases} H_0: \mu_1 = \mu_2 = \mu_3 \\ H_1: \neg(\mu_1 = \mu_2 = \mu_3) \end{cases} \tag{4.35}$$

gdzie:
μ_1, μ_2, μ_3 – wartości oczekiwane skorygowanej o wpływ „SOC" zmiennej „GHQ_suma".

Tabela 4.44. Wyniki porównania wartości oczekiwanych (wzór 4.35) zmiennej „GHQ_suma" w modelu dwuczynnikowej analizy kowariancji

ukl_kraz		Suma kwadratów	*df*	Średni kwadrat	*F*	Istotność	Cząstkowe eta kwadrat	Parametr niecentralności	Moc obserwowana[a]
,0	kontrast	1338,868	2	669,434	12,149	,000	,054	24,298	,995
	błąd	23639,002	429	55,103					
1,0	kontrast	1077,561	2	538,781	9,778	,000	,044	19,556	,983
	błąd	23639,002	429	55,103					

Każde *F* testuje proste efekty „grupa_wieku" w ramach każdej kombinacji poziomów innych przedstawionych efektów; testy te oparte są na liniowo niezależnych porównaniach parami pomiędzy oszacowanymi średnimi brzegowymi.
[a] Obliczone z użyciem alfa = ,05.

Wyniki testów przedstawione w tab. 4.44 możemy rozumieć jako porównywanie wartości oczekiwanych zmiennej „GHQ_suma", skorygowanej o wpływ zmiennej „SOC" w warstwach: ukl_kraz = 0 (osoby wolne od chorób układu krążenia) i ukl_kraz = 1 (osoby ze zdiagnozowaną chorobą układu krążenia) między grupami wiekowymi. Prawdopodobieństwo w obu przypadkach jest mniejsze od 0,0005, cząstkowe eta kwadrat wskazuje na efekt słaby (na podstawie wartości parametru; nic nie wiemy o merytorycznym znaczeniu obserwowanych różnic). A obserwowana moc testu jest bliska jedności i do niczego nie jest w tym przypadku potrzebna. Mogłaby przydawać się przy *p* w teście większym od 0,05, ale wówczas prawie zawsze jest za mała, aby na jej podstawie można było przyjąć hipotezę zerową.

Tabela 4.45. Wyniki porównań parami wartości oczekiwanych zmiennej „GHQ_suma" skorygowanej o wpływ zmiennej kowariancyjnej „SOC" między kategoriami zmiennej „ukl_kraz" w grupach wiekowych

grupa_wieku	(I) ukl_kraz	(J) ukl_kraz	Różnica średnich (I–J)	Błąd standardowy	Istotność[b]	95% przedział ufności dla różnicy[b]	
						dolna granica	górna granica
1	,0	1,0	–1,897	1,412	,180	–4,673	,879
	1,0	,0	1,897	1,412	,180	–,879	4,673
2	,0	1,0	–2,684*	1,128	,018	–4,901	–,466
	1,0	,0	2,684*	1,128	,018	,466	4,901
3	,0	1,0	–4,047*	1,913	,035	–7,806	–,287
	1,0	,0	4,047*	1,913	,035	,287	7,806

Na podstawie estymowanych średnich brzegowych.
* Różnica średnich jest istotna na poziomie ,05.
[b] Poprawka dla porównań wielokrotnych – Bonferroniego.

Tabela 4.46. Wyniki porównania wartości oczekiwanych zmiennej „GHQ_suma" między kategoriami zmiennej „ukl_kraz" w poszczególnych grupach wieku w modelu dwuczynnikowej analizy kowariancji

grupa_wieku		Suma kwadratów	*df*	Średni kwadrat	*F*	Istotność	Cząstkowe eta kwadrat	Parametr niecentralności	Moc obserwowana[a]
1	kontrast	99,418	1	99,418	1,804	,180	,004	1,804	,268
	błąd	23 639,002	429	55,103					
2	kontrast	311,772	1	311,772	5,658	,018	,013	5,658	,660
	błąd	23 639,002	429	55,103					

grupa_wieku		Suma kwadratów	*df*	Średni kwadrat	*F*	Istotność	Cząstkowe eta kwadrat	Parametr niecentralności	Moc obserwowana[a]
3	kontrast	246,566	1	246,566	4,475	,035	,010	4,475	,560
	błąd	23 639,002	429	55,103					

Każde *F* testuje proste efekty ukl_kraz w ramach każdej kombinacji poziomów innych przedstawionych efektów; testy te oparte są na liniowo niezależnych porównaniach parami pomiędzy oszacowanymi średnimi brzegowymi.

[a] Obliczone z użyciem alfa = ,05.

Wyniki przedstawione w tab. 4.46 możemy interpretować jako efekt porównywania skorygowanych o wpływ zmiennej „SOC" wartości oczekiwanych zmiennej „GHQ_suma" między osobami bez zdiagnozowanych chorób układu krążenia (ukl_kraz = 0) i osobami, u których zdiagnozowano jakąś chorobę układu krążenia (ukl_kraz = 1) w poszczególnych grupach wieku.

W najmłodszej grupie wieku (grupa_wieku = 1) $p = 0{,}180$, $\eta_p^2 = 0{,}004$, a obserwowana moc testu jest równa 0,268. Nie ma podstaw do odrzucenia hipotezy zerowej (wniosek z testowania); brak efektu (na podstawie wielkości η_p^2) i obserwowana moc testu nie umożliwia nam przyjęcia hipotezy zerowej o braku różnic między wartościami oczekiwanymi – klasyka gatunku. W dwóch pozostałych grupach wiekowych wartości oczekiwane zmiennej „GHQ_suma" w grupie „zdrowych" i grupie „chorych" są różne, wielkość efektu jest na poziomie umiarkowanym (lecz nie jest to ocena merytoryczna), zaś moc obserwowana równa 0,660 i 0,560 do niczego się nie przydaje.

PRZYKŁAD 4.10

W tym przykładzie przeanalizujemy (w potrzebnym nam zakresie) wyniki eksperymentu toksykologicznego, dla analizy których zastosowano dwuczynnikowy model analizy wariancji z powtarzaniem obserwacji na jednym czynniku. Eksperyment polegał na jednorazowym podaniu szczurom butoksyetanolu w czterech różnych dawkach (każda grupa zwierząt otrzymała inną dawkę) oraz sprawdzaniu w kilku punktach czasowych parametrów krwi. W eksperymencie tym chodziło m.in. o sprawdzenie, czy butoksyetanol ma działanie hemotoksyczne. Czynnik randomizowalny, czyli „grupa", to zmienna przyjmująca cztery wartości – są cztery grupy zwierząt (cztery dawki), zaś czynnik nierandomizowalny, tj. „czynnik1", przyjmuje pięć wartości, odpowiadających pięciu pomiarom w kolejnych punktach czasowych: punkt 0 – pomiar przed narażeniem, a następnie po 4, 11, 18 i 28 dniach po narażeniu (Starek i wsp., 2006). W przykładzie analizowany jest parametr „czerwone krwinki" (RBC). Zmienna „czynnik1" to czas.

```
GLM RBC0 RBC4 RBC11 RBC18 RBC28 BY grupa
/WSFACTOR=czynnik1 5 Polynomial
/METHOD=SSTYPE(3)
/EMMEANS=TABLES(grupa) COMPARE ADJ(BONFERRONI)
/EMMEANS=TABLES(czynnik1) COMPARE ADJ(BONFERRONI)
/EMMEANS=TABLES(grupa*czynnik1) COMPARE(grupa) ADJ(BONFERRONI)
/EMMEANS=TABLES(grupa*czynnik1) COMPARE(czynnik1) ADJ(BONFERRONI)
/PRINT=DESCRIPTIVE ETASQ OPOWER HOMOGENEITY
/CRITERIA=ALPHA(.05)
/WSDESIGN=czynnik1
/DESIGN=grupa.
```

Tabela 4.47. Wyniki testu Mauchly'ego sferyczności macierzy wariancji-kowariancji

Efekt wewnątrz-obiektowy	W Mauchly'ego	Przybliżone chi-kwadrat	*df*	Istotność	Epsilon		
					Greenhouse-Geisser	Huynh-Feldt	dolna granica
czynnik1	,102	32,897	9	,000	,468	,628	,250

Test Mauchly'ego, czyli test sferyczności macierzy wariancji-kowariancji, można traktować jako pewną analogię testu jednorodności wariancji w metodach analizy wariancji. Sferyczność macierzy wariancji-kowariancji pozwala korzystać z pierwszej linii „sferyczność założona" w testach efektów wewnątrzobiektowych. Gdy założenie o sferyczności macierzy nie jest spełnione, powinniśmy korzystać z testów Greehouse-Geissera albo Huynh-Feldta.

Tabela 4.48. Wyniki testów efektów wewnątrzobiektowych w dwuczynnikowej analizie wariancji

Źródło		Typ III sumy kwadratów	*df*	Średni kwadrat	*F*	Istotność	Cząstkowe eta kwadrat	Parametr niecentralności	Moc obserwowana
czynnik1	sferyczność założona	17,170	4	4,293	24,856	,000	,608	99,423	1,000
	Greenhouse-Geisser	17,170	1,873	9,166	24,856	,000	,608	46,561	1,000
	Huynh-Feldt	17,170	2,510	6,839	24,856	,000	,608	62,400	1,000
	dolna granica	17,170	1,000	17,170	24,856	,000	,608	24,856	,997
czynnik1 * grupa	sferyczność założona	22,445	12	1,870	10,830	,000	,670	129,964	1,000
	Greenhouse-Geisser	22,445	5,620	3,994	10,830	,000	,670	60,864	1,000
	Huynh-Feldt	22,445	7,531	2,980	10,830	,000	,670	81,569	1,000
	dolna granica	22,445	3,000	7,482	10,830	,000	,670	32,491	,993

Źródło		Typ III sumy kwadratów	*df*	Średni kwadrat	*F*	Istotność	Cząstkowe eta kwadrat	Parametr niecentralności	Moc obserwowana
Błąd (czynnik1)	Sferyczność założona	11,053	64	,173					
	Greenhouse-Geisser	11,053	29,972	,369					
	Huynh-Feldt	11,053	40,168	,275					
	dolna granica	11,053	16,000	,691					

[a] Obliczone z użyciem alfa = ,05.

Tabela 4.49. Wyniki testu Levene'a równości wariancji zmiennej RBC w poszczególnych punktach czasowych

Zmienna	*F*	*df*1	*df*2	Istotność
RBC0	3,407	3	16	,043
RBC4	1,019	3	16	,410
RBC11	2,897	3	16	,067
RBC18	3,053	3	16	,059
RBC28	1,049	3	16	,398

Tabela 4.50. Wyniki testów efektów międzyobiektowych dla zmiennej „grupa"; porównywanie „średnich" RBC w grupach zwierząt

Źródło	Typ III sumy kwadratów	*df*	Średni kwadrat	*F*	Istotność	Cząstkowe eta kwadrat	Parametr niecentralności	Moc obserwowana
Stała	7675,512	1	7675,512	24 083,816	,000	,999	24 083,816	1,000
Grupa	41,111	3	13,704	42,998	,000	,890	128,995	1,000
Błąd	5,099	16	,319					

W tej konkretnej analizie obliczane są cząstkowe η^2, obserwowane moce testów oraz przedstawione są wyniki testowania odpowiednich hipotez. Zarówno w przypadku oceny efektów wewnątrzobiektowych, jak i międzyobiektowych odpowiednie prawdopodobieństwa są mniejsze od 0,0005. Świadczy to tylko o tym, że wartości oczekiwane badanego parametru (losowej zmiennej ciągłej, tutaj RBC) nie są jednakowe we wszystkich punktach czasowych ani we wszystkich

porównywanych grupach zwierząt oraz że istnieje interakcja między oboma czynnikami. Stosunkowo duże wartości cząstkowych η^2 pozwalają przypuszczać, że zależności między czynnikiem randomizowalnym i badanym parametrem oraz między czynnikiem nierandomizowalnym i badanym parametrem są dość silne (oczywiście nie jest to ocena merytoryczna, a jedynie ocena na podstawie miernika). Co możemy wnioskować na temat interakcji na podstawie wartości cząstkowej $\eta^2 = 0{,}670$? Nie wiem. Co może oznaczać stwierdzenie, że efekt interakcji jest silny? O czym mówimy, wspominając o efekcie interakcji? Czy stwierdzenie faktu, że zależność między czynnikami i badanym parametrem jest dość silna lub nawet bardzo silna, pogłębia w czymś naszą wiedzę o relacjach między efektem narażenia i podaną dawką, o charakterze zmienności w czasie badanego parametru czy o charakterze interakcji między czynnikami? Nie, dalej trzeba prowadzić analizę, stosując testy porównań wielokrotnych dla efektów głównych, oceniać efekty proste i używać testów porównań wielokrotnych dla efektów prostych. Przyjrzyjmy się kolejnym fragmentom wydruku wyników analizy.

Tabela 4.51. Statystyki opisowe zmiennej RBC w porównywanych grupach zwierząt w kolejnych punktach czasowych eksperymentu

grupa	czynnik1	Średnia	Błąd standardowy	95% przedział ufności	
				dolna granica	górna granica
0	1	9,620	,158	9,284	9,956
	2	9,980	,340	9,259	10,701
	3	9,040	,129	8,766	9,314
	4	9,500	,154	9,174	9,826
	5	9,580	,145	9,273	9,887
1	1	9,340	,158	9,004	9,676
	2	9,400	,340	8,679	10,121
	3	8,500	,129	8,226	8,774
	4	9,060	,154	8,734	9,386
	5	9,220	,145	8,913	9,527
2	1	9,240	,158	8,904	9,576
	2	7,860	,340	7,139	8,581
	3	8,340	,129	8,066	8,614
	4	8,580	,154	8,254	8,906
	5	8,840	,145	8,533	9,147

grupa	czynnik1	Średnia	Błąd standardowy	95% przedział ufności	
				dolna granica	górna granica
3	1	9,560	,158	9,224	9,896
	2	6,240	,340	5,519	6,961
	3	7,320	,129	7,046	7,594
	4	7,880	,154	7,554	8,206
	5	8,120	,145	7,813	8,427

Tabela 4.52. Wyniki testów porównań parami między „średnimi" w poszczególnych grupach w kolejnych punktach czasowych – czynnik1 (fragment tabeli)

czyn-nik1	(I) grupa	(J) grupa	Różnica średnich (I–J)	Błąd standardowy	Istotność	95% przedział ufności dla różnicy	
						dolna granica	górna granica
1	0	1	,280	,224	1,000	–,393	,953
		2	,380	,224	,654	–,293	1,053
		3	,060	,224	1,000	–,613	,733
	1	0	–,280	,224	1,000	–,953	,393
		2	,100	,224	1,000	–,573	,773
		3	–,220	,224	1,000	–,893	,453
	2	0	–,380	,224	,654	–1,053	,293
		1	–,100	,224	1,000	–,773	,573
		3	–,320	,224	1,000	–,993	,353
	3	0	–,060	,224	1,000	–,733	,613
		1	,220	,224	1,000	–,453	,893
		2	,320	,224	1,000	–,353	,993
2	0	1	,580	,481	1,000	–,867	2,027
		2	2,120	,481	,003	,673	3,567
		3	3,740	,481	,000	2,293	5,187
	1	0	–,580	,481	1,000	–2,027	,867
		2	1,540	,481	,033	,093	2,987
		3	3,160	,481	,000	1,713	4,607
	2	0	–2,120	,481	,003	–3,567	–,673
		1	–1,540	,481	,033	–2,987	–,093
		3	1,620	,481	,023	,173	3,067

Tabela 4.53. Porównanie wartości oczekiwanych zmiennej RBC między grupami zwierząt w poszczególnych punktach czasowych

czynnik1		Suma kwadratów	*df*	Średni kwadrat	*F*	Istotność	Cząstkowe eta kwadrat	Parametr niecentralności	Moc obserwowana
1	kontrast	,484	3	,161	1,288	,312	,195	3,864	,279
	błąd	2,004	16	,125					
2	kontrast	42,250	3	14,083	24,355	,000	,820	73,065	1,000
	błąd	9,252	16	,578					
3	kontrast	7,748	3	2,583	31,023	,000	,853	93,069	1,000
	błąd	1,332	16	,083					
4	kontrast	7,222	3	2,407	20,400	,000	,793	61,199	1,000
	błąd	1,888	16	,118					
5	kontrast	5,852	3	1,951	18,622	,000	,777	55,866	1,000
	błąd	1,676	16	,105					

Interpretacja wyników zawartych w tej tabeli jest analogiczna, jak w przykładzie 4.9 (tab. 4.44).

Tabela 4.54. Wyniki testów porównań parami między „średnimi" w poszczególnych punktach czasowych (czynnik1) w kolejnych grupach zwierząt (fragment tabeli)

grupa	(I) czynnik1	(J) czynnik1	Różnica średnich (I–J)	Błąd standardowy	Istotność	95% przedział ufności dla różnicy	
						dolna granica	górna granica
0	1	2	–,360	,354	1,000	–1,512	,792
		3	,580	,190	,076	–,038	1,198
		4	,120	,199	1,000	–,528	,768
		5	,040	,200	1,000	–,610	,690
	2	1	,360	,354	1,000	–,792	1,512
		3	,940	,334	,126	–,148	2,028
		4	,480	,379	1,000	–,754	1,714
		5	,400	,371	1,000	–,805	1,605

grupa	(I) czynnik1	(J) czynnik1	Różnica średnich (I–J)	Błąd standardowy	Istotność	95% przedział ufności dla różnicy	
						dolna granica	górna granica
	3	1	–,580	,190	,076	–1,198	,038
		2	–,940	,334	,126	–2,028	,148
		4	–,460	,145	,060	–,932	,012
		5	–,540	,155	,031	–1,044	–,036
	4	1	–,120	,199	1,000	–,768	,528
		2	–,480	,379	1,000	–1,714	,754
		3	,460	,145	,060	–,012	,932
		5	–,080	,106	1,000	–,426	,266
	5	1	–,040	,200	1,000	–,690	,610
		2	–,400	,371	1,000	–1,605	,805
		3	,540	,155	,031	,036	1,044
		4	,080	,106	1,000	–,266	,426

Tabela 4.55. Porównanie wartości oczekiwanych zmiennej RBC między punktami czasowymi w poszczególnych grupach zwierząt

grupa		Wartość	*F*	*df* hipotezy	*df* błędu	Istotność	Cząstkowe eta kwadrat	Parametr niecentralności	Moc obserwowana
0	Ślad Pillai	,597	4,807	4,000	13,000	,013	,597	19,227	,849
	Lambda Wilksa	,403	4,807	4,000	13,000	,013	,597	19,227	,849
	Ślad Hotellinga	1,479	4,807	4,000	13,000	,013	,597	19,227	,849
	Największy pierwiastek Roya	1,479	4,807	4,000	13,000	,013	,597	19,227	,849
1	Ślad Pillai	,692	7,310	4,000	13,000	,003	,692	29,241	,965
	Lambda Wilksa	,308	7,310	4,000	13,000	,003	,692	29,241	,965
	Ślad Hotellinga	2,249	7,310	4,000	13,000	,003	,692	29,241	,965
	Największy pierwiastek Roya	2,249	7,310	4,000	13,000	,003	,692	29,241	,965

Tabela 4.54 (cd.)

grupa		Wartość	*F*	*df* hipotezy	*df* błędu	Istotność	Cząstkowe eta kwadrat	Parametr niecentralności	Moc obserwowana
2	Ślad Pillai	,680	6,892	4,000	13,000	,003	,680	27,569	,955
	Lambda Wilksa	,320	6,892	4,000	13,000	,003	,680	27,569	,955
	Ślad Hotellinga	2,121	6,892	4,000	13,000	,003	,680	27,569	,955
	Największy pierwiastek Roya	2,121	6,892	4,000	13,000	,003	,680	27,569	,955
3	Ślad Pillai	,915	35,082	4,000	13,000	,000	,915	140,329	1,000
	Lambda Wilksa	,085	35,082	4,000	13,000	,000	,915	140,329	1,000
	Ślad Hotellinga	10,795	35,082	4,000	13,000	,000	,915	140,329	1,000
	Największy pierwiastek Roya	10,795	35,082	4,000	13,000	,000	,915	140,329	1,000

Interpretacja wyników zawartych w tej tabeli jest analogiczna jak w przykładzie 4.9 (tab. 4.46).

W tabelach porównań parami (wyniki testów porównań wielokrotnych) brak zarówno mierników wielkości efektów, jak i obserwowanej mocy testu. Parametry te są w tab. 4.44, 4.46, 4.53 i 4.55, jednak relacja między prawdopodobieństwem w teście (nieszczęśliwie nazywanym w SPSS „istotnością") a obserwowaną mocą testu jest ciągle taka sama: małe prawdopodobieństwo w teście to duża wartość obserwowanej mocy testu i na odwrót – duża wartość prawdopodobieństwa to mała wartość obserwowanej mocy. Korzyść z obliczania obserwowanej mocy testu jest zatem żadna. Miernik wielkości efektu, cząstkowe eta kwadrat też nie ułatwia nam podjęcia decyzji, gdyż nie niesie żadnej informacji merytorycznej.

4.3.4. 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. Jak zauważają Tabachnik i Fidell (2007), proponowano już wiele miar w regresji logistycznej, które są odpowiednikami R^2 z wielozmiennowych modeli regresji liniowej. Jednak nie istnieje powszechnie akceptowany bezpośredni odpowiednik R^2 z regresji OLS (*Ordinary Least*

Square). Jest to spowodowane interpretacją R^2, czyli „procentem wyjaśnianej wariancji", ale wariancja zmiennej dwustanowej albo dyskretnej zmiennej zależnej jest uzależniona od rozkładu częstości tej zmiennej. To powoduje, że współczynniki R^2 szacowane w modelach regresji logistycznej dla różnych zmiennych wynikowych z badania nie mogą być porównywane bezpośrednio, jak również problematyczne jest porównywanie R^2 w modelu regresji logistycznej i R^2 w modelu regresji liniowej (OLS). Wszystkie one (R^2 w modelach regresji logistycznej) mogą być traktowane jako aproksymacja OLS R^2, a nie jako rzeczywisty procent wyjaśnionej wariancji. Jednak należy być świadomym, że wielu badaczy wykazuje jedynie marginalne zainteresowanie tymi namiastkami *R*-kwadrat, uważając, iż preferowaną miarą wielkości efektu jest – dyskutowany poniżej – wskaźnik klasyfikacji. Zauważmy, że R^2-podobne miary, nazywane najczęściej pseudo-R^2, przedstawione niżej, nie są testami jakości dopasowania, lecz raczej próbą mierzenia siły zależności. Na nieszczęście, pseudomiary R^2 odzwierciedlają i mieszają siłę efektu z jakością dopasowania. Na przykład dla małych próbek miary pseudo-R^2 mogą przyjmować wysokie wartości, gdy jakość dopasowania w teście ilorazu wiarygodności była nieakceptowalna. W SPSS-ie obliczane są dwa pseudo-R^2 mierniki: Coxa i Snella, Nagelkerke'a (Garson, 2012).

Zamiast mówić o zmiennej zależnej w regresji logistycznej, poprawniej będzie mówić o zmiennej wynikowej z badania. Zmienną zależną w tych modelach, przez analogię do modeli regresji liniowej, będzie prawdopodobieństwo zdarzenia określonego przez wartość zmiennej wynikowej z badania.

W modelach regresji liniowej współczynniki regresji zazwyczaj, choć nie zawsze, są szacowane metodą najmniejszych kwadratów. Tego typu modele nazywane bywają modelami OLS (*Ordinary Least Squares*). Natomiast w modelach regresji logistycznej współczynniki regresji szacowane są metodą największej wiarygodności. Czym różnią się metody szacowania wartości współczynników regresji: najmniejszych kwadratów i największej wiarygodności?

W modelach regresji logistycznej (a dokładniej: binarnej regresji logistycznej, gdyż do takiej się ograniczę) zmienna wynikowa z badania jest zmienną dwustanową:

$$\mathrm{Y} = \begin{cases} 0 & \text{badane zdarzenie nie występuje} \\ 1 & \text{badane zdarzenie występuje} \end{cases} \tag{4.36}$$

Zależność między prawdopodobieństwem badanego zdarzenia a zmiennymi objaśniającymi opisywana jest wzorem:

$$P(Y=1) = \frac{1}{1+exp(-(B_0+B_1 \cdot X_1+B_2 \cdot X_2+\cdots+B_n \cdot X_n))} \tag{4.37}$$

albo

$$P(Y=1) = \frac{exp(B_0+B_1 \cdot X_1+B_2 \cdot X_2+\cdots+B_n \cdot X_n)}{1+exp(B_0+B_1 \cdot X_1+B_2 \cdot X_2+\cdots+B_n \cdot X_n)} \tag{4.38}$$

gdzie:
$X_1, X_2, \ldots, X_n$ – są zmiennymi objaśniającymi, w regresji logistycznej nazywanymi czynnikami ryzyka; mogą one być zmiennymi ciągłymi albo dyskretnymi.

Wzór opisujący zależność między prawdopodobieństwem badanego zjawiska i czynnikami ryzyka można też spotkać w postaci:

$$P(Y = 1|X) = \frac{1}{1+exp(-(B_0+B_1 \cdot X_1+B_2 \cdot X_2+\cdots+B_n \cdot X_n))} \tag{4.39}$$

gdzie:
$X = (X_1, X_2, \ldots, X_n)^T$; litera T w wykładniku oznacza transpozycję wektora X, tzn. wektor ten jest wektorem pionowym.

Oznaczmy przez X wektor n czynników ryzyka $(X_1, X_2, \ldots, X_n)$. Prawdopodobieństwo warunkowe w modelu regresji logistycznej oznaczmy przez $\pi(X)$:

$$P(Y = 1|X) = \pi(X) \tag{4.40}$$

Przyjmując oznaczenie:

$$g(X) = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \cdots + \beta_n \cdot X_n \tag{4.41}$$

możemy napisać:

$$\pi(X) = \frac{e^{g(X)}}{1+e^{g(X)}} \tag{4.42}$$

Logarytm funkcji wiarygodności dla modelu jednozmiennowego, czyli gdy:

$$\pi(X) = \frac{e^{\beta_0+\beta_1 \cdot X}}{1+e^{\beta_0+\beta_1 \cdot X}} \tag{4.43}$$

wyraża się wzorem:

$$l(\beta) = ln[L(\beta)] = \sum_{i=1}^{k}\{y_i \cdot ln[\pi(x_i)] + (1 - y_i) \cdot ln[1 - \pi(x_i)]\} \tag{4.44}$$

a sama funkcja wiarygodności wzorem:

$$L(\beta) = exp[l(\beta)] \tag{4.45}$$

gdzie:
(x_i, y_i), $i = 1, 2, \ldots, k$ są zaobserwowanymi wartościami zmiennych X i Y w próbce.

Funkcja wiarygodności $L(\beta)$ jest funkcją wielu zmiennych opisującą pewną wielowymiarową powierzchnię scharakteryzowaną przez współczynniki β (wzory 4.44 i 4.45). $\beta = (\beta_0, \beta_1, \beta_2, \ldots, \beta_n)$ jest wektorem. Na powierzchni $L(\beta)$ poszukujemy maksimum, wartości maksimum traktujemy jako oszacowania wartości

parametrów β_0, β_1, β_2, …, β_n. Nazwa metody szacowania współczynników regresji, metoda największej wiarygodności, staje się oczywista – oszacowane współczynniki to maksimum funkcji wiarygodności. Metoda największej wiarygodności szacowania współczynników regresji jest metodą iteracyjną, tzn. w następnym kroku wykorzystywane są wyniki uzyskane w kroku poprzednim. Proces ten jest zatrzymywany, gdy różnica wartości między kolejnymi krokami nie przekracza ustalonej małej wartości, np. 0,001.

Przyjmijmy następujące oznaczenia, aby móc przedstawić wzory dla R^2 zaimplementowane w SPSS i dla najpopularniejszego pseudo-R^2, czyli R_L-kwadrat:

- 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:

$$R_{CS}^2 = 1 - \left(\frac{L_0}{L_F}\right)^{\frac{2}{n}} \tag{4.46}$$

zaś współczynnik Nagelkerke'a wzorem:

$$R_N^2 = \frac{1-(L_0/L_F)^{\frac{2}{n}}}{1-(L_0)^{\frac{2}{n}}} \tag{4.47}$$

R^2 Coxa i Snella próbuje imitować w zakresie interpretacji współczynnik determinacji w wielozmiennowym modelu regresji liniowej. Jednak maksimum tego miernika może być (i zazwyczaj jest) mniejsze od jedności, co doskonale utrudnia jego interpretację.

R^2 Nagelkerke'a jest modyfikacją współczynnika Coxa i Snella, zapewniającą jego zmienność od 0 do 1. R^2 Nagelkerke'a zazwyczaj będzie większy od miernika Coxa i Snella, ale na ogół bywa mniejszy niż R^2 w modelu regresji liniowej (Nagelkerke, 1991).

Różne „mutacje" pseudo-R^2, zarówno te przedstawione powyżej, jak i inne, omawiane są w pracach takich autorów, jak: 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):

$$R_L^2 = 1 - \frac{ln(L_F)}{ln(L_0)} \tag{4.48}$$

Menard (2000) napisał, że „po pierwsze i najważniejsze, R_L^2 ma najbardziej intuicyjnie uzasadnioną interpretację jako proporcjonalna redukcja miary błędu porównywalna z klasycznym R^2". Jednakże stosując R_L^2 nie jesteśmy pewni wpływu predyktorów (czynników ryzyka) na rezultat. Co na przykład oznacza $R_L^2 = 0{,}10$

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 oceny jakości 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.

Tabachnik i Fidell (2007: 463) sugerują, powołując się na pracę Chinn (2000), że można przekształcić iloraz szans do współczynnika Cohena *d*, który z kolei może być przekształcony w η^2:

$$d = ln(OR)\ /1{,}81$$
$$\eta^2 = \frac{d^2}{d^2+4} \tag{4.49}$$

Pomijając magiczne działanie dzielenia przez 1,81 oraz sposób obliczania η^2, to czy zaproponowany sposób oceny wielkości efektu znajduje zastosowanie zarówno w przypadku ciągłych czynników ryzyka, jak i dyskretnych? Dla dyskretnych czynników ryzyka otrzymujemy ilorazy szans (ORs – *Odds Ratios*) dla każdej wartości tego czynnika w stosunku 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.

Inna proponowana miara wielkości efektu to procent poprawnych klasyfikacji, przedstawiony w tabeli klasyfikacji. Problem w tym, że badacze nie potrafią się zgodzić, jaka wielkość tego miernika świadczy o dużej czy małej wielkości efektu. Z drugiej strony, dla stosunkowo rzadkich zdarzeń miernik ten jest całkowicie fałszywy.

PRZYKŁAD 4.11

Poszukujemy czynników ryzyka wystąpienia chorób układu krążenia wśród pracowników służb mundurowych (Dudek, 2007). Jako potencjalne czynniki ryzyka wystąpienia chorób układu krążenia potraktowaliśmy następujące zmienne:

$$\text{grupa\_wieku} = \begin{cases} 1 & \text{do 35 lat} \\ 2 & 36-45 \text{ lat} \\ 3 & \text{46 albo więcej lat} \end{cases}$$

zmienna „grupa_wieku" to zmienna dyskretna zbudowana ze zmiennej „wiek" (wiek wyrażony w latach);

$$\text{SOC\_fa} = \begin{cases} 1 & \text{do 100} \\ 2 & 101-149 \\ 3 & \text{150 albo więcej} \end{cases}$$

zmienna „SOC_fa" powstała ze zmiennej ciągłej „SOC";

$$\text{palenie3} = \begin{cases} 0 & \text{nigdy niepalący} \\ 1 & \text{aktualni palacze} \\ 2 & \text{byli palacze} \end{cases}$$

oraz zmienne: „czyn_wyp", „bier_wyp", „brak_snu", „HDL", „LDL" i „TG".

LOGISTIC REGRESSION VARIABLES **ukl_kraz**
/METHOD=BSTEP(WALD) grupa_wieku SOC_fa czyn_wyp bier_wyp brak_snu palenie3 HDL LDL TG
/CONTRAST (grupa_wieku)=Simple(1)
/CONTRAST (SOC_fa)=Simple(1)
/CONTRAST (palenie3)=Simple(1)
/PRINT=ITER(1) CI(95)
/CRITERIA=PIN(0.05) POUT(0.051) ITERATE(20) CUT(0.5).

Tabela 4.56. Blok początkowy w modelowaniu regresji logistycznej; model jedynie ze stałą, w którym obliczana jest wartość funkcji wiarygodności dla takiego modelu[a, b, c]

Iteracja		–2 logarytm wiarygodności	Współczynniki
			Stała
Krok 0	1	485,264	–,785
	2	485,098	–,829
	3	485,098	–,829

[a] Stała została włączona do modelu.
[b] Początkowa wartość –2 logarytm wiarygodności: 485,098.
[c] Estymacja została zakończona na iteracji o numerze 3, ponieważ oszacowania parametrów zmieniły się o mniej niż ,001.

W tab. 4.56 dla bloku początkowego pokazane są wartości –2log funkcji wiarygodności (L_0) modelu zawierającego tylko stałą.

Tabela 4.57. Iteracyjny proces szacowania współczynników modelu; zastosowano metodę eliminacji wstecznej z kryterium Walda[a, b, c, d, e]

Iteracja		–2 logarytm wiarygodności	Współczynniki												
			stała	grupa_wieku(1)	grupa_wieku(2)	SOC_fa(1)	SOC_fa(2)	czyn_wyp	bier_wyp	brak_snu	palenie3(1)	palenie3(2)	HDL	LDL	TG
Krok 1	1	428,432	–,993	,412	,894	–1,031	–1,202	–,035	–,006	,157	–,211	–,039	–,009	,008	,000
	2	421,245	–1,180	,601	1,124	–1,085	–1,331	–,063	–,011	,211	–,331	–,062	–,014	,011	,001
	3	420,988	–1,205	,639	1,165	–1,086	–1,345	–,072	–,012	,221	–,359	–,068	–,015	,012	,001
	4	420,987	–1,206	,640	1,167	–1,086	–1,345	–,073	–,012	,221	–,360	–,069	–,015	,012	,001
	5	420,987	–1,206	,640	1,167	–1,086	–1,345	–,073	–,012	,221	–,360	–,069	–,015	,012	,001
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Krok 6	1	438,658	–1,946	,475	1,054			–,042		,170				,008	
	2	433,705	–2,448	,662	1,279			–,070		,219				,011	
	3	433,581	–2,512	,689	1,307			–,078		,225				,012	
	4	433,581	–2,513	,690	1,308			–,078		,225				,012	
	5	433,581	–2,513	,690	1,308			–,078		,225				,012	

[a] Metoda: selekcja wsteczna (Wald).
[b] Stała została włączona do modelu.
[c] Początkowa wartość –2 logarytm wiarygodności: 485,098 (wartość z tabeli: Przebieg iteracji w Bloku 0).
[d] Estymacja została zakończona na iteracji o numerze 5, ponieważ oszacowania parametrów zmieniły się o mniej niż ,001.
[e] Estymacja została zakończona na iteracji o numerze 4, ponieważ oszacowania parametrów zmieniły się o mniej niż ,001.

W tab. 4.57 pokazane są m.in. wartości –2log funkcji wiarygodności (L_F) modelu zawierającego wszystkie potencjalne czynniki ryzyka (krok 1) oraz wszystkie statystycznie istotne czynniki ryzyka (krok 6). Nie pokazywałem etapów pośrednich, czyli kroków 2–5. Obliczone wartości funkcji wiarygodności posłużą do obliczenia pseudo-R^2.

Tabela 4.58. Wartości: –2 logarytm wiarygodności i pseudo-R^2 Coxa i Snella oraz Nagelkerke'a w kolejnych krokach tworzenia modelu końcowego (w tym przypadku zawierającego tylko istotne statystycznie czynniki ryzyka chorób układu krążenia)

Krok	–2 logarytm wiarygodności	*R*-kwadrat Coxa i Snella	*R*-kwadrat Nagelkerke'a
1	420,987[a]	,150	,212
2	421,145[a]	,149	,211
3	422,992[b]	,145	,206
4	427,157[b]	,136	,193
5	429,808[b]	,131	,185
6	433,581[a]	,122	,173

[a] Estymacja została zakończona na iteracji o numerze 5, ponieważ oszacowania parametrów zmieniły się o mniej niż ,001.

[b] Estymacja została zakończona na iteracji o numerze 4, ponieważ oszacowania parametrów zmieniły się o mniej niż ,001.

Prezentowane w tab. 4.58 wartości pseudo-R^2 Coxa i Snella oraz Nagelkerke'a stosunkowo łatwo obliczyć samodzielnie. W tab. 4.56 i 4.57 podane są wartości –2×logarytm(wartość funkcji wiarygodności). Możemy więc policzyć wartości funkcji wiarygodności oraz podstawić je do wzorów (4.46) i (4.47).

Tabela 4.59. Wartości logarytmu funkcji wiarygodności w bloku 0 (tylko ze stałą) (tab. 4.56) i w bloku 1 (zawierającym czynniki ryzyka) (tab. 4.57)

Wyrażenie	Wartość wyrażenia	
	Blok 0 (tab. 4.56)	Blok 1 (tab. 4.57)
–2ln(funkcji wiarygodności)	485,098	433,581
ln(funkcji wiarygodności)	–242,549	–216,7905
wartość funkcji wiarygodności	4,5952E–106*	7,06451E–95*

* Zapis 4,5952E–106 oznacza $4{,}5952 \cdot 10^{-106}$ i analogicznie 7,0645E–95 to $7{,}0645 \cdot 10^{-95}$.

Pseudo-R^2 Coxa i Snella jest równe:

$$\text{pseudo } R^2 \text{ Coxa i Snella} = 1 - \left(\frac{L_0}{L_F}\right)^{\frac{2}{n}} = 1 - \left(\frac{4{,}5925 \cdot 10^{-106}}{7{,}06451 \cdot 10^{-95}}\right)^{\frac{2}{395}} = 0{,}122276 \tag{4.50}$$

Co dla oceny wielkości efektu, w przypadku modelu regresji logistycznej, wynika z pokazanych w tabeli wartości pseudo-R^2? Brak efektu, efekt słaby, średni czy jeszcze inny?

Tabela 4.60. Tabela klasyfikacji osób „zdrowych" i „chorych" na choroby układu krążenia przy wykorzystaniu zbudowanego modelu regresji logistycznej[a]

Kroki analizy	Obserwowane		Przewidywane		
			ukl_kraz		procent poprawnych klasyfikacji
			,0	1,0	
Krok 1	ukl_kraz	,0	250	25	90,9
		1,0	81	39	32,5
	procent ogółem				73,2
Krok 6	ukl_kraz	,0	254	21	92,4
		1,0	89	31	25,8
	procent ogółem				72,2

[a] Punktem podziału jest ,500.

A co możemy powiedzieć o wielkości efektu na podstawie procentu poprawnych klasyfikacji? Procent poprawnych klasyfikacji osób wolnych od chorób układu krążenia to 92,4% – bardzo dużo, ale procent poprawnych klasyfikacji osób ze zdiagnozowanymi chorobami układu krążenia to tylko 25,8%, czyli raczej mało. Ogólny procent poprawnych klasyfikacji łącznie wśród osób, u których nie zdiagnozowano chorób układu krążenia oraz tych ze zdiagnozowanymi chorobami układu krążenia to 72,2%. Prawie 3/4 osób, które znalazły się w próbie, zaklasyfikowanych zostało prawidłowo za pomocą utworzonego modelu. Jednak tak duży procent poprawnie zaklasyfikowanych osób jest tutaj efektem stosunkowo mało licznej grupy osób, u których zdiagnozowano choroby układu krążenia. Jeszcze gorzej będzie to wyglądało w przypadku np. chorób układu oddechowego, gdyż osób ze zdiagnozowanymi chorobami układu oddechowego jest jeszcze mniej.

Tabela 4.61. Wyniki szacowania współczynników regresji i ilorazów szans (Exp(*B*) = OR (*Odds Ratio*)) w ostatnim kroku tworzenia modelu regresji logistycznej dla zmiennej wynikowej z badania: choroby układu krążenia

Istotne czynniki ryzyka		*B*	Błąd standardowy	Wald	*df*	Istotność	Exp(*B*)	95% przedział ufności dla EXP(*B*)	
								dolna granica	górna granica
Krok 6[a]	grupa_wieku			13,933	2	,001			
	grupa_wieku(1)	,690	,268	6,643	1	,010	1,994	1,180	3,369
	grupa_wieku(2)	1,308	,362	13,031	1	,000	3,697	1,818	7,520
	czyn_wyp	–,078	,033	5,535	1	,019	,925	,867	,987
	brak_snu	,225	,065	11,872	1	,001	1,252	1,102	1,423
	LDL	,012	,004	8,285	1	,004	1,012	1,004	1,020
	stała	–2,513	,566	19,684	1	,000	,081		

[a] Zmienne wprowadzone w kroku 1: grupa_wieku, SOC_fa, czyn_wyp, bier_wyp, brak_snu, palenie3, HDL, LDL, TG.

Tabela 4.62. Fragment tab. 4.61 z dodatkowymi kolumnami, w których znalazły się oszacowania wielkości efektu

Istotne czynniki ryzyka		*B*	Błąd standardowy	Wald	*df*	*p*	OR	Wielkość efektu	
								d	η^2
Krok 6	grupa_wieku			13,933	2	,001			
	grupa_wieku(1)	,690	,268	6,643	1	,010	1,994	0,381	0,092
	grupa_wieku(2)	1,308	,362	13,031	1	,000	3,697	0,722	0,160
	czyn_wyp	–,078	,033	5,535	1	,019	,925	–0,043	–0,011
	brak_snu	,225	,065	11,872	1	,001	1,252	0,124	0,031
	LDL	,012	,004	8,285	1	,004	1,012	0,007	0,002
	stała	–2,513	,566	19,684	1	,000	,081	–1,389	–0,234

Tabela 4.63. Porównanie interpretacji wielkości efektu na podstawie wartości mierników d i η^2

Miernik d*		Miernik η^2**		
przedziały	interpretacja wielkości efektu	przedziały według Cohena (1988)	przedziały według Sinka i Mvududu (2007)	interpretacja wielkości efektu
$d < 0{,}2$	brak efektu	$\eta^2 < 0{,}01$	$\eta^2 < 0{,}01$	brak efektu
$0{,}2 \leq d < 0{,}5$	efekt słaby	$0{,}01 \leq \eta^2 < 0{,}09$	$0{,}01 \leq \eta^2 < 0{,}06$	efekt słaby
$0{,}5 \leq d < 0{,}8$	efekt średni	$0{,}09 \leq \eta^2 < 0{,}25$	$0{,}06 \leq \eta^2 < 0{,}14$	efekt średni (umiarkowany)
$d \geq 0{,}8$	efekt silny	$\eta^2 \geq 0{,}25$	$\eta^2 \geq 0{,}14$	efekt silny (duży)

* Szczegóły w podrozdziale 4.3.1.
** Szczegóły w podrozdziale 4.3.3.

Rodzi się tutaj pewna wątpliwość. Miernik d wielkości efektu mierzył ten efekt w zagadnieniu porównywania dwóch średnich, zaś miernik η^2 w metodach analizy wariancji, gdy porównywanych było kilka średnich. Czy znaczenie tych mierników jest takie samo w zagadnieniach regresji logistycznej? Czy możemy porównywać wartości tych mierników (oczywiście porównujemy wartości mierników d między sobą i mierników η^2 między sobą) dla czynnika ryzyka dyskretnego (jakim jest grupa wieku) i ciągłego (czynny wypoczynek, brak snu i LDL)?

Przyjrzyjmy się decyzjom podejmowanym na podstawie wartości mierników wielkości efektu (znak w tym momencie nie jest ważny, gdyż mówimy tylko o sile, a nie o kierunku zależności). Dla zmiennej „czyn_wyp" (czynny wypoczynek) $d = 0{,}043$, zaś $\eta^2 = 0{,}011$. Decyzja na podstawie miernika d: brak efektu, zaś według miernika η^2 efekt jest słaby. Natomiast prawdopodobieństwo w teście hipotez:

$$\begin{cases} H_0: \mathrm{OR} = 1 \\ H_1: \mathrm{OR} \neq 1 \end{cases} \tag{4.51}$$

jest równe 0,019 i jest mniejsze od 0,05, czyli od strony statystycznej OR jest różne od 1. Czy możemy na podstawie wartości d mówić, że nie ma efektu, albo na podstawie η^2, że efekt niby jest, ale słaby? Co to znaczy, że jest słaby? Zmienna „czyn_wyp" to liczba godzin poświęcanych tygodniowo na czynny wypoczynek wymagający bardzo dużego (regularny trening, udział w zawodach) lub dużego (bieganie, siłownia, pływanie, gra w piłkę, biegi narciarskie, taniec dyskotekowy) wysiłku fizycznego. Wraz ze zwiększeniem liczby godzin czynnego wypoczynku (o dużej intensywności) o 1, ryzyko chorób układu krążenia zmienia się 0,925 razy. To przy jednej godzinie, natomiast przy wzroście liczby godzin poświęcanych na czynny wypoczynek w ciągu tygodnia o 5, ryzyko będzie mniejsze 0,677 razy, bez mała spada o połowę!

Zmienna „brak_snu": ile razy w ciągu tygodnia zdarza się P. spać mniej niż P. powinien/powinna? Decyzje na podstawie mierników wielkości efektu są takie same, jak w przypadku zmiennej opisującej czynny wypoczynek – brak efektu albo efekt słaby. Ale jeśli człowiek regularnie nie będzie dosypiał swojej fizjologicznej normy (wartość zmiennej „brak_snu" = 7), to ryzyko powstania u niego chorób układu krążenia będzie 4,82 ($1{,}252^7 = 4{,}822$) razy większe niż u człowieka wysypiającego się zgodnie ze swoim zapotrzebowaniem na sen. Czyżby to rzeczywiście był brak efektu?!

Chciałbym jeszcze zwrócić uwagę Czytelnika na zmienną „grupa_wieku". Jeśli dyskretny czynnik ryzyka nie jest kodowany jako: 0, 1, …, to program SPSS przekodowuje wartości takiej zmiennej, aby najniższa kategoria była oznaczona jako 0. Dlatego też „grupa_wieku(1)" oznacza grupę wiekową 36–45 lat, a „grupa_wieku(2)" to osoby mające 46 lat albo starsze. Grupa wiekowa do 35 lat, jako zadeklarowana grupa referencyjna, nie jest wyświetlana w tabeli wyników; przyjmuje się dla niej (przyjmuje, a nie oblicza, szacuje), iż OR jest równe 1. OR dla grupy 36–45 lat jest równe 1,994; dla grupy 46 lat lub więcej OR = 3,697. Dla grupy „grupa_wieku(2)" $d = 0{,}722$, $\eta^2 = 0{,}160$. Według d jest to efekt średni. Taki sam jest w oparciu o η^2 według Cohena, natomiast według Sinka i Mvududu jest to już efekt silny. A jakie znaczenie mają te ryzyka z lekarskiego punktu widzenia? Z rachunków to nie wynika.

PRZYKŁAD 4.12

W tym przykładzie przedstawiam wybrane fragmenty modelowania funkcji regresji logistycznej w przypadku chorób układu oddechowego.

```
LOGISTIC REGRESSION VARIABLES ukl_odde
/METHOD=BSTEP(WALD) grupa_wieku SOC_fa czyn_wyp bier_wyp brak_snu palenie3 HDL LDL TG
/CONTRAST (grupa_wieku)=Simple(1)
/CONTRAST (SOC_fa)=Simple(1)
/CONTRAST (palenie3)=Simple(1)
/PRINT=GOODFIT CI(95)
/CRITERIA=PIN(0.05) POUT(0.051) ITERATE(20) CUT(0.5).
```

Tabela 4.64. Wartości: –2 logarytm wiarygodności i pseudo-R^2 Coxa i Snella oraz Nagelkerke'a w kolejnych krokach tworzenia modelu końcowego (w tym przypadku zawierającego tylko istotne statystycznie czynniki ryzyka chorób układu oddechowego)

Krok	–2 logarytm wiarygodności	*R*-kwadrat Coxa i Snella	*R*-kwadrat Nagelkerke'a
1	89,507[a]	,061	,242
⋮	⋮	⋮	⋮
9	104,576[b]	,024	,097

[a] Estymacja została zakończona na iteracji o numerze 20 z powodu osiągnięcia maksymalnej liczby iteracji; nie jest możliwe uzyskanie ostatecznego rozwiązania.

[b] Estymacja została zakończona na iteracji o numerze 8, ponieważ oszacowania parametrów zmieniły się o mniej niż ,001.

Tabela 4.65. Tabela klasyfikacji osób „zdrowych" i „chorych" na choroby układu oddechowego przy wykorzystaniu zbudowanego modelu regresji logistycznej[a]

Kroki analizy	Obserwowane		Przewidywane		
			ukl_odde		procent poprawnych klasyfikacji
			,0	1,0	
Krok 1	ukl_odde	,0	382	0	100,0
		1,0	13	0	,0
	procent ogółem				96,7
⋮	⋮	⋮	⋮	⋮	⋮
Krok 9	ukl_odde	,0	382	0	100,0
		1,0	13	0	,0
	procent ogółem				96,7

[a] Punktem podziału jest ,500.

Ogólny procent poprawnie zaklasyfikowanych osób to 96,7, mimo iż żadna z osób, u których zdiagnozowano choroby układu oddechowego, nie została poprawnie zaklasyfikowana na podstawie zbudowanego modelu. Z jednej strony, wyraźnie widać, że miernik wielkości efektu w postaci ogólnego procentu poprawnie zaklasyfikowanych osób nie jest żadnym miernikiem. Z drugiej strony, ten wysoki odsetek poprawnych klasyfikacji kłóci się z wartościami współczynników pseudo-R^2 Coxa i Snella oraz Nagelkerke'a, które są bardzo niewielkie. Wydaje mi się, że spójne to nie jest.

Tabela 4.66. Wyniki szacowania współczynników regresji i ilorazów szans (Exp(*B*) = OR (*Odds Ratio*)) w ostatnim kroku tworzenia modelu regresji logistycznej dla zmiennej wynikowej z badania: choroby układu oddechowego

Istotne wyniki ryzyka		*B*	Błąd standardowy	Wald	*df*	Istotność	Exp(*B*)	95% przedział ufności dla EXP(*B*)	
								dolna granica	górna granica
Krok 9[a]	grupa_wieku			6,124	2	,047			
	grupa_wieku(1)	2,380	1,050	5,134	1	,023	10,801	1,379	84,618
	grupa_wieku(2)	1,123	1,423	,623	1	,430	3,075	,189	50,026
	stała	–3,926	,486	65,376	1	,000	,020		

[a] Zmienne wprowadzone w kroku 1: grupa_wieku, SOC_fa, czyn_wyp, bier_wyp, brak_snu, palenie3, HDL, LDL, TG.

Po ostatnim kroku w modelu pozostała tylko jedna zmienna, w istotny sposób związana z ryzykiem chorób układu oddechowego – „grupa_wieku". Pozostałe zmienne okazały się nieistotne. Wyjaśnienie, dlaczego ryzyko chorób układu oddechowego w najstarszej grupie wieku jest mniejsze od ryzyka w grupie młodszej (36–45 lat) wymaga konsultacji specjalisty.

4.4. Merytoryczne znaczenie obserwowanych różnic i wielkość efektu

W badaniach medycznych, biologicznych, a także psychologicznych, nawet jeśli uzyskamy tzw. istotność statystyczną badanej zależności (najczęściej będzie to dotyczyło różnic między wartościami oczekiwanymi), to musimy zastanowić się, czy uzyskane różnice mają dla nas, jako badaczy, jakieś znaczenie merytoryczne. Niekiedy okazuje się, że „istotna statystycznie" różnica jest mniejsza od dokładności pomiaru badanego parametru. I fakt „istotności statystycznej" nic nam nie daje. Interpretacja „merytorycznego znaczenia" uzyskanych zależności jest określana przez badacza i zmienia się od badania do badania. Artykuły tego typu ukazują się w czasopismach „branżowych", np. „Drug and Alcohol Review" (Miller, Manuel, 2008), „Journal of Rehabilitation Medicine" (Donoghue i wsp., 2009), „Respiratory Medicine" (De Kleijn i wsp., 2011), „European Journal of Vascular and Endovascular Surgery" (Frans i wsp., 2014), „Journal of Behavioral Health Services and Research" (Eisen i wsp., 2007). Poniżej kilka przykładów dotyczących merytorycznego znaczenia ocenianych parametrów.

Frans i wsp. (2014) zauważają, że nawet jeśli dla pacjentów z chorobami naczyń obwodowych rejestrowane są wyniki badań diagnostycznych, to z punktu widzenia pacjenta, a często także lekarza ważniejsze są zmiany w ocenie jakości życia jako miary efektywności leczenia. I „istotność statystyczna" zmian parametrów diagnostycznych nie przekłada się na „istotność" oceny jakości życia. Ten problem nie jest bezpośrednio związany z merytoryczną oceną znaczenia parametrów diagnostycznych, lecz z używaniem dwóch różnych kryteriów oceny stanu zdrowia pacjentów. Jak zauważają autorzy tej pracy, należałoby określić „minimalną wartość różnic (zmian)" parametrów diagnostycznych, które by określały kliniczną wartość zmian w ocenie jakości życia, mniej przywiązując się do istotności statystycznej.

W artykule De Kleijna i wsp. (2011) porównywane są wyniki dwóch metod „statystycznych" do oszacowania minimalnej klinicznie ważnej zmiany (MCID – *Minimal Clinically Important Difference*) u pacjentów chorych na sarkoidozę. Wartość MCID określano, używając metody wartości granicznej (*anchor-based method*) i metod wykorzystujących rozkład (*distribution-based methods*).

Dla obu tych sposobów szacowania MCID uzyskano wartość około 4 na skali oceny wyczerpania (FAS – *Fatigue Assessment Scale*) i wartość tę przyjęto jako mającą znaczenie medyczne: pacjenci, którzy w dwóch badaniach uzyskiwali przyrost wskaźnika nieco większy od 4, byli uznawani za osoby, których stan zdrowia się poprawił, natomiast ci, których zmiana wskaźnika była mniejsza niż –4, traktowane były jako osoby, których stan zdrowia się pogorszył. Zmiana wskaźnika w przedziale od –4 do 4 traktowana była jako nieistotna i osoby te uznawano za osoby z brakiem poprawy, ale bez pogorszenia. Przy wyznaczaniu MCID posługiwano się również prawdopodobieństwem w testach dotyczących współczynnika korelacji, lecz nie prawdopodobieństwo miało tu podstawowe znaczenie.

Jak już Czytelnik mógł zauważyć na przykładzie publikacji Frans i wsp. (2014), także w innych badaniach pojawia się pewna dwoistość wyników. Mierzone i oceniane statystycznie są wartości pewnych parametrów, zaś np. efekt leczenia oceniany jest inaczej, chociaż mierzone parametry mają w tym swój udział. Mimo że ten problem nie polega na różnej ocenie statystycznej i merytorycznej badanego parametru, to też ma ogromne znaczenie przy ocenie wyników analizy statystycznej. Wielu badaczy, np. cytowani już Chmura-Kraemer i Kupfer (2006), uważają, iż ocena wielkości efektu może służyć do oceny ważności merytorycznej uzyskanego wyniku statystycznego. Co więcej – sugerują, że ocena wielkości efektu jest w tej sytuacji niezbędna[7]. Ale z taką opinią trudno się zgodzić w świetle poprzednich rozważań na temat wielkości efektów w różnych modelach statystycznych.

4.5. Wielkość efektu dla metod nieparametrycznych

Dotychczasowe rozważania dotyczące oceny wielkości efektów dotyczyły metod parametrycznych. W przypadku konieczności zastosowania metody nieparametrycznej rodzą się dwa problemy. Pierwszy to fakt, że metody parametryczne i nieparametryczne nie prowadzą do uzyskania równoważnych rozwiązań. Przyjrzyjmy się jednoczynnikowej metodzie analizy wariancji. W wersji parametrycznej mamy do czynienia z następującym zagadnieniem testowania:

$$\begin{cases} H_0: \mu_1 = \mu_2 = \cdots = \mu_k \\ H_1: \neg(\mu_1 = \mu_2 = \cdots = \mu_k) \end{cases} \tag{4.52}$$

gdzie w hipotezie zerowej mamy zapisaną równość wartości oczekiwanych w analizowanych grupach, zaś w hipotezie alternatywnej, iż wartości oczekiwane w grupach nie są jednakowe. W „wersji" nieparametrycznej:

7 „To judge the clinical significance of a statistically significant finding, an effect size is needed".

$$\begin{cases} H_0: F_1 = F_2 = \cdots = F_k \\ H_1: \neg(F_1 = F_2 = \cdots = F_k) \end{cases} \tag{4.53}$$

w hipotezach mamy do czynienia z rozkładami prawdopodobieństwa.

Zarówno w podejściu parametrycznym, jak i nieparametrycznym możemy podjąć jedną z dwóch decyzji: odrzucić hipotezę zerową i przyjąć alternatywną albo uznać, że nie ma podstaw do odrzucenia hipotezy zerowej. W przypadku nieodrzucenia hipotezy zerowej w wersji nieparametrycznej praktycy (badacze stosujący metody statystyczne) uznają, że porównywane rozkłady prawdopodobieństwa nie różnią się. Skoro rozkłady się nie różnią to znaczy, że wszystkie charakteryzujące je parametry, w tym wartości oczekiwane, też będą jednakowe, co jest zapisane w hipotezie zerowej w wersji parametrycznej. Możemy w tym momencie uznać, iż decyzje w wersji parametrycznej i nieparametrycznej są równoważne.

Jeśli odrzucimy hipotezę zerową w wersji nieparametrycznej i przyjmiemy nieparametryczną hipotezę alternatywną, to będziemy musieli zgodzić się, że rozkłady prawdopodobieństwa analizowanej cechy w poszczególnych grupach nie są jednakowe. Ale nie jesteśmy w stanie powiedzieć, w czym te rozkłady nie są jednakowe. Jakie prawo pozwala nam przypuszczać, że wartości oczekiwane nie będą jednakowe? I w tym przypadku nie ma żadnej równoważności w podejściu parametrycznym i nieparametrycznym.

Drugi problem, jaki pojawia się przy stosowaniu metod nieparametrycznych jest taki, iż nie możemy korzystać z mierników oceny wielkości efektu wykorzystywanych w podejściu parametrycznym. Potrzebne są specyficzne mierniki dla rozwiązań nieparametrycznych. Jednak w większości programów do analizy statystycznej danych takie mierniki nie są zaimplementowane. Pojawiają się wszakże pewne wzory umożliwiające obliczenie wartości tych mierników, które ich autorzy nazywają wielkością efektu oraz proponowane są pewne przedziały tych mierników, umożliwiające nazwanie efektu dużym, małym czy średnim.

Oczywiście, w materiałach tworzonych przez statystyków nie ma nawet wzmianki o czymś takim, jak wielkość efektu (np. Lehmann, 2006; Wasserman, 2006), ale o tym zagadnieniu pisałem już wcześniej.

Wuensch (2015) w materiałach dydaktycznych ze statystyki przedstawia, za innymi autorami, dwa wzory, które mają służyć jako ocena wielkości efektu w przypadku dwóch prób niezależnych (test Manna-Whitneya) albo zależnych (test Wilcoxona), nie podając jednak granic pozwalających na klasyfikację tych wielkości efektów. I tak, Gray i Kinnear (2012) proponują:

$$\frac{2(M_1 - M_2)}{n_1 + n_2} \tag{4.54}$$

gdzie M_1 i M_2 są średnimi sumy rang w każdej z prób, a n_1 i n_2 to liczebności tych prób. Zaś Kerby (2014) proponuje wzór:

$$1 - \frac{2U}{n_1 \cdot n_2} \tag{4.55}$$

gdzie *U* jest wartością statystyki Manna-Whitheya.

U Fielda (2009) znajdujemy natomiast wzór dla dwóch prób niezależnych porównywanych testem Manna-Whitneya:

$$r = \frac{z}{\sqrt{N}} \tag{4.56}$$

gdzie *z* jest pochodną statystyki będącej podstawą testu obliczaną w SPSS, a dla *r* mamy już granice pozwalające uznać wielkość efektu jako słaby, średni lub duży. W przypadku liczby grup większej niż dwie stosowany jest test Kruskala-Wallisa i nie znalazłem dla niego oceny wielkości efektu. Field używa tutaj mierników, pochodnych testu Manna-Whitneya, dla poszczególnych par porównań, stawiając pod znakiem zapytania sens używania testów porównań wielokrotnych.

4.6. Krótkie podsumowanie rozdziału o ocenie wielkości efektu

W tym rozdziale zwróciłem uwagę na pewne słabości klasycznej teorii statystyki, choć raczej powinniśmy mówić o klasycznych teoriach, ale także na konsekwencje wymieszania dwóch teorii statystycznych: 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 spośród 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). Istnieją jednak argumenty przeciwko ocenie wielkości efektów (Denis, 2003). Co prawda, argumenty, na które powołuje się Denis, pochodzą od innych autorów. Sam Denis uważa, że ocena wielkości efektu przynosi więcej korzyści niż strat.

Na podstawie przejrzanego piśmiennictwa zaobserwowałem następującą prawidłowość. W tekstach pisanych przez matematyków i statystyków nie pojawiają się metody oceny wielkości efektów – takie pojęcie najczęściej w ogóle w tych pracach nie występuje. Natomiast w tekstach pisanych przez psychologów – oczywiście tam, gdzie jest sens wykorzystywania metod statystycznych – zawsze

występuje wielkość efektu. Jest to wcześniejsze zderzenie podejścia do statystyki teoretyków statystyków i badaczy ją wykorzystujących.

W przypadku psychologów dochodzi jeszcze presja instytucjonalna APA (American Psychological Association). W wytycznych na temat publikacji znajduje się wymaganie, aby podawać wielkość i kierunek efektu (APA, 2010)[8].

Rozważania zawarte w obecnym rozdziale można by podsumować następująco: teoria testowania hipotez Neymana-Pearsona ma pewne luki, które utrudniają jej stosowanie, ale rozwiązania dotyczące oceny wielkości efektu wcale tych luk nie wypełniają. Co więcej, teoria Neymana-Pearsona ma przyzwoite uwarunkowania teoretyczne i teoretyczną nadbudowę konkretnych metod, a ocena wielkości efektu, wykorzystując elementy probabilistyczne, jedynie stwarza pozory poprawności metodologicznej.

8 „For inferential statistical tests (e.g. *t*, *F*, and 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. When point estimates (e.g. sample means or regression coefficients) are provided, always include an associated measure of variability (precision), with an indication of the specific measure used (e.g. the standard error)".

Rozdział 5. O innych podejściach do wnioskowania statystycznego

5.1. Wprowadzenie

Przedstawiając toczącą się przynajmniej od czterdziestu lat dyskusję o potrzebie zmiany paradygmatu statystyki, musimy być w pełni świadomi, czego ów paradygmat dotyczy. Czy będziemy mówili o paradygmacie statystyki matematycznej w pełnej ogólności, czyli paradygmacie odnoszącym się do testowania hipotez statystycznych, metod estymacji, problemów podejmowania decyzji itd., czy może zawęzimy to pojęcie. W rzeczywistości najwięcej emocji wśród badaczy stosujących metody statystyki matematycznej do opracowywania wyników badań budzą zagadnienia związane z testowaniem hipotez statystycznych. Zarówno teoria Fishera, jak i teoria Neymana-Pearsona testowania hipotez statystycznych są niesatysfakcjonujące praktycznie dla wszystkich badaczy prowadzących badania w naukach społecznych, humanistycznych, medycznych, biologicznych, rolniczych i innych. Trudno znaleźć dziedzinę nauki, której przedstawiciele nie mieliby zastrzeżeń do rezultatów uzyskiwanych przy wykorzystywaniu obecnych metod testowania hipotez. Dlatego też mówiąc o istniejących i oczekiwanych paradygmatach czy paradygmatach statystyki, będę miał na myśli głównie paradygmaty testowania hipotez statystycznych.

Wracając do pojęcia paradygmatu, warto przypomnieć efekt ewolucji pojęcia paradygmatu w nauce w rozważaniach Kuhna. Nastąpiło wyraźne przeniesienie pojęcia paradygmatu z osiągnięć naukowych na grupę badaczy. Pojęcie paradygmatu w tym sensie będzie mało adekwatne do analizowanego problemu testowania hipotez statystycznych. Wydaje mi się, że znacznie bardziej adekwatne sformułowanie to użyty przez Billa Thompsona (2007) zwrot „statystyczne modele indukcji". Ponieważ pojęcie „model" jest zbyt konkretne, np. model regresji liniowej, nieliniowej, model analizy wariancji itp., więc nie bardzo nadaje się jako zamiennik pojęcia „paradygmat". W dalszych rozważaniach będę używał sformułowania „teoria testowania" albo „paradygmat testowania" hipotez.

W rozdziale pierwszym przedstawiłem podstawy teorii testowania zaproponowane przez Fishera i Neymana-Pearsona. Teorie te można potraktować jako dwa różne paradygmaty testowania hipotez statystycznych – i tak są one traktowane przez wielu autorów. Na przykład Lillestøl (2014) w swoim opracowaniu wymienia aż pięć paradygmatów:

- wczesne wnioskowanie bayesowskie i jego odrodzenie się (*revival*);
- wnioskowanie fisherowskie;
- wnioskowanie Neymana-Pearsona;
- neobayesowskie wnioskowanie;
- wnioskowanie wiarygodnościowe.

Aitkin (2011), rozumiejąc paradygmat w sensie zaproponowanym przez Kuhna jako rozłączne społeczności uczonych stosujących dane metody, wyróżnia praktycznie dwa paradygmaty: bayesowski i częstościowy, dopuszczając istnienie innych (np. wiarygodnościowego), którymi jednak nie zajmuje się w swoim opracowaniu.

Z kolei w książce *Philosophy of Statistics* pod redakcją Bandyopadhyaya i Forstera (2011) znajdujemy cztery paradygmaty: klasyczny paradygmat statystyki, paradygmat bayesowski, paradygmat wiarygodnościowy i paradygmat Akaike. Nie wchodząc w spory filozoficzne, przedstawię pokrótce dwie najczęściej wykorzystywane teorie statystyczne.

5.2. Metody bayesowskie (paradygmat bayesowski)

Istnieje wiele opracowań omawiających podstawy i zastosowania metod bayesowskich (np. Carlin, Louis, 2000; Congdon, 2002; Spiegelhalter i wsp., 2004). Wszystkie rozważania wychodzą od twierdzenia Bayesa o prawdopodobieństwie *a posteriori*. Większość modeli bayesowskich składa się z dwóch etapów: z wiarygodnościowej specyfikacji rozkładu zmiennej Y zależnej od pewnego parametru θ:

$$Y|\theta \sim f(Y|\theta) \tag{5.1}$$

a następnie porachowania odpowiedniego prawdopodobieństwa warunkowego – wzory (5.2) i (5.3).

Wzór (5.1) informuje nas, że rozkład prawdopodobieństwa zmiennej Y, zależnej od parametru θ, opisany jest funkcją gęstości $f(y)$, która również zależy od parametru θ. Mówię tutaj o parametrze θ, lecz w podejściu bayesowskim θ jest zmienną losową, o pewnym rozkładzie prawdopodobieństwa:

$$\theta \sim \pi(\theta) \tag{5.2}$$

W najprostszych analizach bayesowskich przyjmuje się, że rozkład prawdopodobieństwa parametru θ jest rozkładem znanym, określonym. Wówczas możemy obliczyć prawdopodobieństwo warunkowe przy zaobserwowanej wartości zmiennej *Y*:

$$p(\theta|y) = \frac{f(y|\theta)\pi(\theta)}{m(y)} \tag{5.3}$$

gdzie:

$$m(y) = \int f(y|\theta)\pi(\theta)d\theta \tag{5.4}$$

jest gęstością brzegową dla wartości *y*.

Ponieważ dla większości badaczy w naukach społecznych z powyższych wzorów nic nie wynika, przedstawię efekty zastosowania podejścia bayesowskiego na przykładzie modelowania ryzyka choroby nowotworowej jako efektu narażenia inhalacyjnego na benzo(a)piren – jeden z wielopierścieniowych węglowodorów aromatycznych (Szymczak, 1999). Jest to węglowodór powstający podczas niepełnego spalania substancji organicznych, np. podczas grillowania potraw czy wypalania traw.

PRZYKŁAD 5.1

W tym przykładzie pokażę sposób modelowania ryzyka choroby nowotworowej u ludzi jako skutku narażenia na benzo(a)piren, wychodząc jednak z modelu eksperymentalnego na zwierzętach (w czasie opracowywania tego zagadnienia brakowało odpowiedniego modelu epidemiologicznego). Dane będące podstawą modelowania pochodzą z pracy Collinsa i wsp. (1991).

Tabela 5.1. Wyniki eksperymentu na chomikach

Stężenie* (mg/m^3)	Czas ekspozycji (tygodnie)	Łączny czas badań (tygodnie)	Liczba osobników z guzem**	Liczba wszystkich osobników
0	–	96,4	0	27
2,2	95,2	95,2	0	27
9,9	96,4	95,4	9	26
46,5	59,2	59,5	13	25

* Zwierzęta narażane były na benzo(a)piren inhalacyjnie przez 4,5 godziny dziennie przez pierwsze 10 dni i przez 3 godzinny dziennie w pozostałych dniach, a ekspozycja trwała 7 dni w tygodniu.

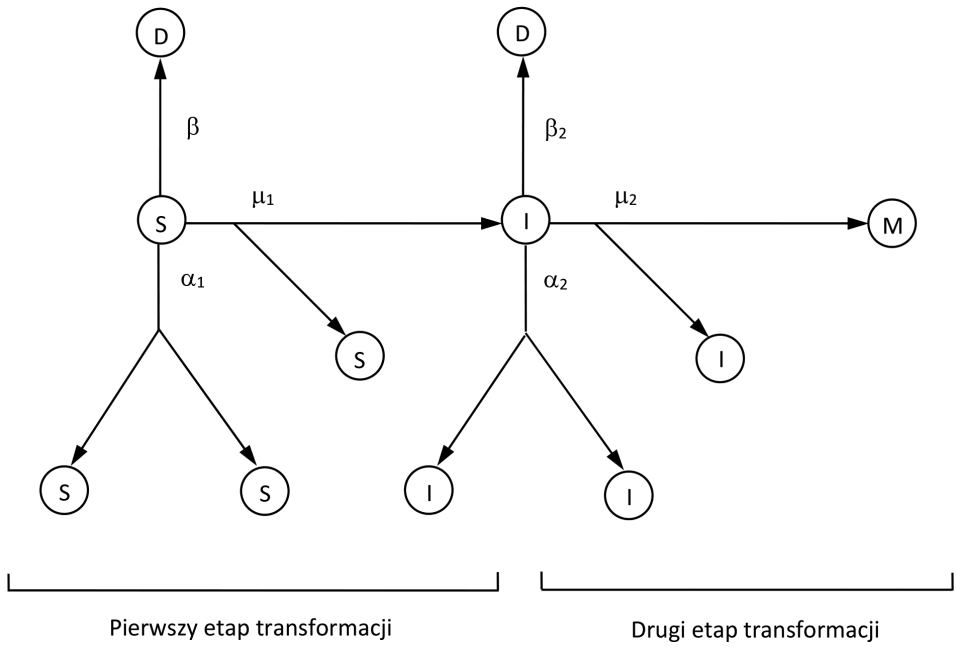
** U narażanych zwierząt obserwowano guzy w jamie nosa, krtani, tchawicy, gardle, przełyku i bezgruczołowej części żołądka; nie zaobserwowano raka płuca. Zaobserwowano następujące guzy: brodawczaki, polipy brodawkowate i raki komórek łuskowatych.

Do obliczenia średniej dawki dla okresu całego życia zwierząt przyjęto następujące parametry allometryczne: dzienne zużycie powietrza przez chomika równe 0,037 m^3/dobę (US EPA, 1984), masę ciała równą 0,12 kg (US EPA, 1988) i przeciętny okres życia chomika 2 lata. Uzyskane średnie dawki, będące podstawą oszacowania zależności dawka-odpowiedź, przedstawiono w tab. 5.2. W tabeli tej nie zamieszczono danych dla stężenia 46,5 mg/m^3, gdyż zwierzęta z ostatniej grupy były narażane znacznie krócej niż te z dwóch pozostałych grup, zatem dane ostatniej grupy nie zostały uwzględnione przy budowie zależności dawka-odpowiedź.

Tabela 5.2. Średnie dawki dla okresu całego życia dla chomika

Stężenie* (mg/m^3)	Czas ekspozycji (tygodnie)	Łączny czas badań (tygodnie)	Dawka całkowita (mg)	Średnia dawka dla okresu całego życia (mg/(kg·dzień)
0	–	96,4	0	0
2,2	95,2	95,2	6,84	0,0781
9,9	96,4	95,4	29,92	0,3416

Do budowy zależności dawka-odpowiedź zostanie wykorzystany najprostszy z wielostopniowych (dwustopniowy) modeli kancerogenezy (Moolgavkar, 1986; Portier i wsp., 1993):



Rycina 5.1. Schemat dwustopniowego modelu kancerogenezy

Oznaczenia na rysunku: *S* – prawidłowa komórka macierzysta, *I* – komórka pośrednia (zainicjowana), *D* – komórka martwa albo komórka, która uległa zróżnicowaniu, *M* – komórka nowotworowa, α_1 – wskaźnik (na komórkę na rok) podziału komórkowego prawidłowych komórek, β_1 – wskaźnik (na komórkę na rok) śmierci albo różnicowania się komórek prawidłowych, μ_1 – wskaźnik (na komórkę na rok) podziału komórki na jedną prawidłową i jedną komórkę pośrednią; α_2, β_2, μ_2 oznaczają odpowiednie wskaźniki dla komórki pośredniej. Funkcja dawka-odpowiedź w tym modelu, nieco uproszczonym w stosunku do przedstawionego na rys. 5.1, ma następującą postać:

$$P(d) = 1 - exp\left\{-\sum_{i=0}^{k} q_i d^i\right\} \; q_i \geq 0; \; i = 0, 1, 2, \ldots, k \tag{5.5}$$

co można zapisać w innej postaci jako:

$$P(d) = 1 - exp\left\{-\sum_{i=0}^{k} q_i d^i\right\} \; q_i \geq 0; \; i = 0, 1, 2, \ldots, k \tag{5.6}$$

gdzie *d* oznacza średnią dawkę dla okresu całego życia wyrażoną w $\frac{\text{mg}}{\text{kg·dzień}}$, zaś q_0, …, q_k są parametrami, których wartości szacuje się metodą największej wiarygodności na podstawie danych empirycznych, natomiast *k* jest liczbą pośrednich etapów biologicznych, przez które przechodzi zdrowa komórka na swej drodze do postaci nowotworowej.

W modelu dwustopniowym funkcja dawka-odpowiedź opisana jest wzorem:

$$P(d) = 1 - exp[-(q_0 + q_1 \cdot d + q_2 \cdot d^2)] \tag{5.7}$$

Wykorzystując dane zawarte w tab. 5.1 i 5.2 oraz stosując metodę największej wiarygodności szacowania współczynników modelu, uzyskujemy funkcję dawka-odpowiedź dla chomika:

$$P(d) = 1 - exp(-3{,}411 \cdot d^2) \tag{5.8}$$

gdzie *d* oznacza średnią dawkę benzo(a)pirenu dla chomika dla okresu całego życia ($\frac{\text{mg}}{\text{kg·dzień}}$), zaś 3,411 jest tutaj parametrem. Aby móc wykorzystać tę funkcję dla oceny ryzyka dla ludzi, musimy przeliczyć dawkę dla człowieka na dawkę dla chomika. Do takiego przeliczenia wykorzystywane są parametry przedstawione w tab. 5.3.

Tabela 5.3. Wartości parametrów wykorzystane podczas budowy dwustopniowego modelu ryzyka choroby nowotworowej jako skutku narażenia na benzo(a)piren (skrót: b(a)p)

Parametr	Wartość parametru wykorzystana w modelowaniu ryzyka
Wskaźnik wentylacji[a] w czasie zmiany roboczej (m^3/8 godz.)	10
Dobowy wskaźnik wentylacji (m^3/24 godz.)	20
Liczba dni pracy w roku	240
Średni czas trwania życia (lata)	70
Średnia masa człowieka (kg)	70
Pomierzone stężenie b(a)p (mg/m^3)	0,0001 (1/20 NDS)
Liczba lat narażenia	3
Średnia dawka „całożyciowa" dla człowieka (mg/(kg·dzień))	$4{,}03 \times 10^{-7}$
Średnia masa chomika (kg)	0,12
Średnia dawka „całożyciowa" dla chomika (mg/(kg·dzień))	$3{,}36 \times 10^{-6}$
Wartość ryzyka dla człowieka (jest to już rezultat modelowania ryzyka)	$3{,}86 \times 10^{-11}$ (4 na 100 mld)

[a] Wskaźnik wentylacji to ilość powietrza zużywanego przez człowieka w określonej jednostce czasu.

Wykorzystując następujący wzór na przekształcenie stężenia benzo(a)pirenu w narażeniu ludzi na dawkę dla chomika (*Calculating Cancer Risk…*, 1995):

$$d = C_z \cdot \frac{10\ \mathrm{m}^3}{20\ \mathrm{m}^3} \cdot \frac{240}{365} \cdot \frac{\text{l. lat naraż.}}{70\ \text{lat}} \cdot \frac{20\ \mathrm{m}^3}{70\ \mathrm{kg}} \cdot \left(\frac{70\ \mathrm{kg}}{0{,}12\ \mathrm{kg}}\right)^{1/3} \tag{5.9}$$

gdzie C_z oznacza stężenie benzo(a)pirenu w środowisku (pracy albo komunalnym), a współczynnik $(70/0{,}12)^{1/3}$ to tzw. współczynnik konwersji, uwzględniający różnicę masy człowieka i zwierzęcia (w tym przypadku chomika), uzyskujemy wartość ryzyka dla człowieka równą $3{,}86 \cdot 10^{-11}$. Jest to ryzyko zachorowania na chorobę nowotworową w wyniku narażenia na benzo(a)piren w stężeniu 0,0001 mg/m^3 przez okres 3 lat w warunkach narażenia zawodowego.

Powyższe oszacowanie ryzyka zostało przeprowadzone przy założeniu, że wszystkie występujące we wzorach parametry są stałymi i mają określoną wartość. Część z nich możemy jednak potraktować jako zmienne losowe (podejście ba-yesowskie) o znanym rozkładzie prawdopodobieństwa.

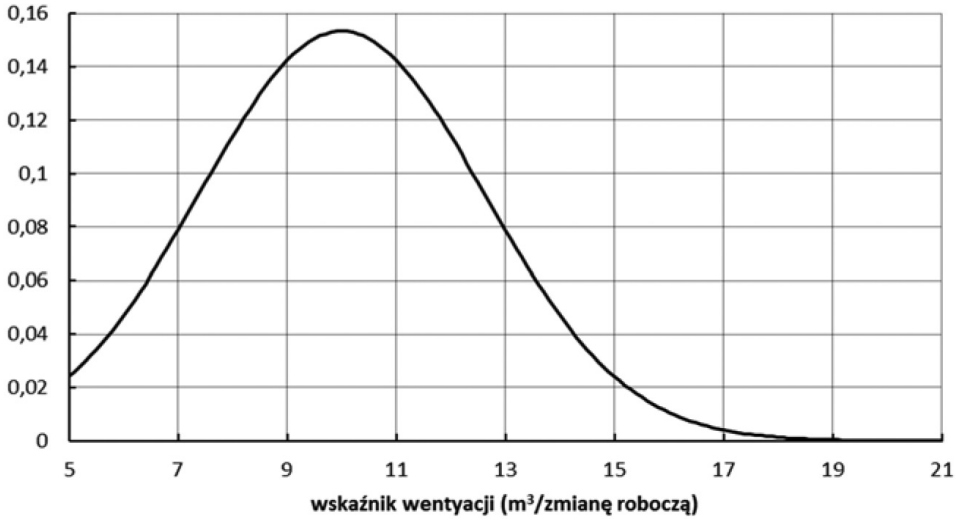
I oczywiście ma to sens. Dlaczego mamy zakładać, że człowiek waży 70 kg – może lepiej przyjąć, że masa człowieka to pewna zmienna losowa? Analogiczna sytuacja dotyczy innych parametrów. Problem, jaki się w tym momencie pojawia, to określenie (przyjęcie) postaci rozkładu prawdopodobieństwa tych zmiennych.

Tabela 5.4. Proponowane rozkłady prawdopodobieństwa parametrów potraktowanych jako zmienne losowe

Parametr w modelu – zmienna	Wartość	Rozkład
Wskaźnik wentylacji (m^3/8 godz.)	10	normalny obcięty
Liczba dni pracy w roku	240	trójkątny
Średni czas trwania życia (lata)	70	normalny obcięty
Średnia masa człowieka (kg)	70	normalny obcięty
Pomierzone stężenie b(a)p (C_Z) (mg/m^3)	0,0001	wartość dokładna
Liczba lat narażenia	3	wartość dokładna
Średnia masa chomika (kg)	0,12	normalny obcięty

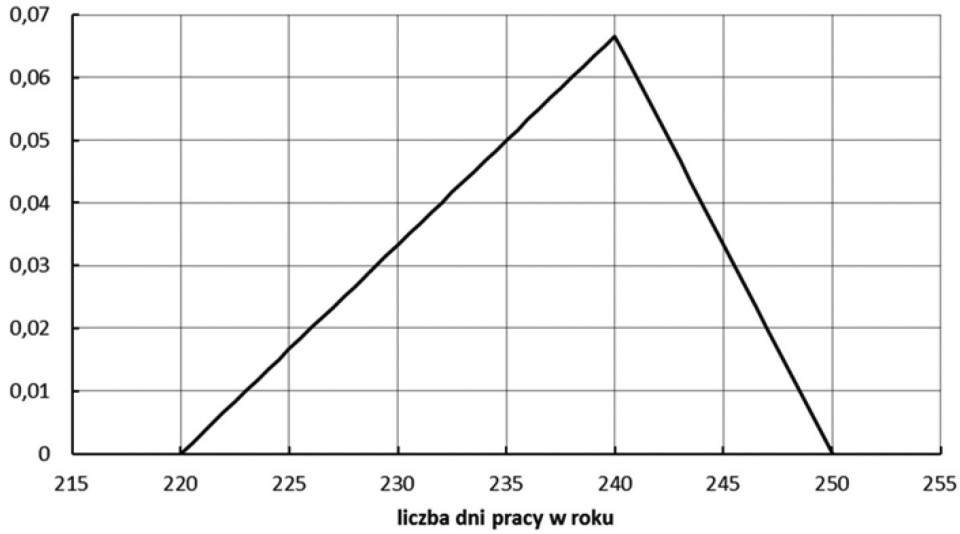
Poniżej przedstawiam dokładniejsze charakterystyki rozkładów prawdopodobieństwa pięciu parametrów.

1. Normalny rozkład obcięty o parametrach: średnia 10,00, odchylenie standardowe 2,60, dobrany zakres od 5,00 do 20,60:



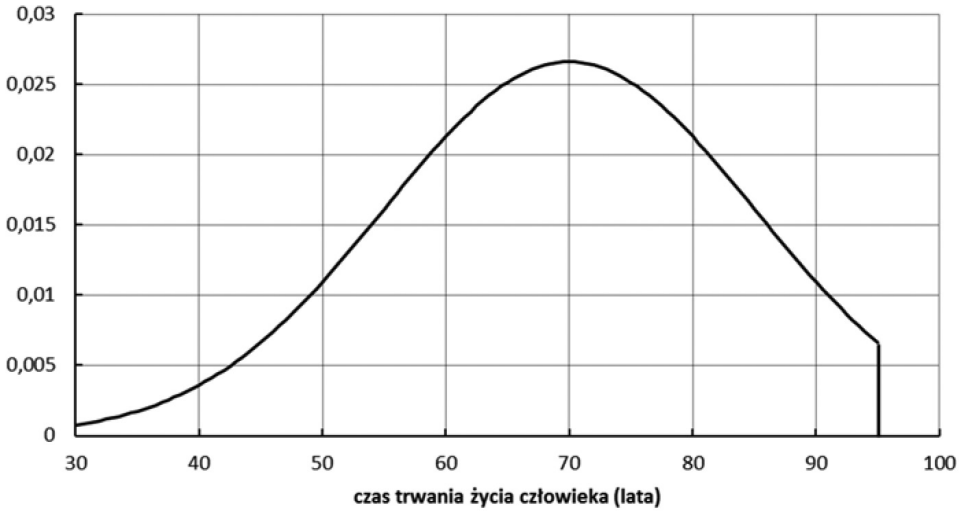
Rycina 5.2. Wykres funkcji gęstości rozkładu prawdopodobieństwa zmiennej: wskaźnik wentylacji w okresie zmiany roboczej

2. Rozkład trójkątny o parametrach: wartość minimalna 220,00, najprawdopodobniejsza 240,00, wartość maksymalna 250,00, dobrany zakres od 220,00 do 250,00:



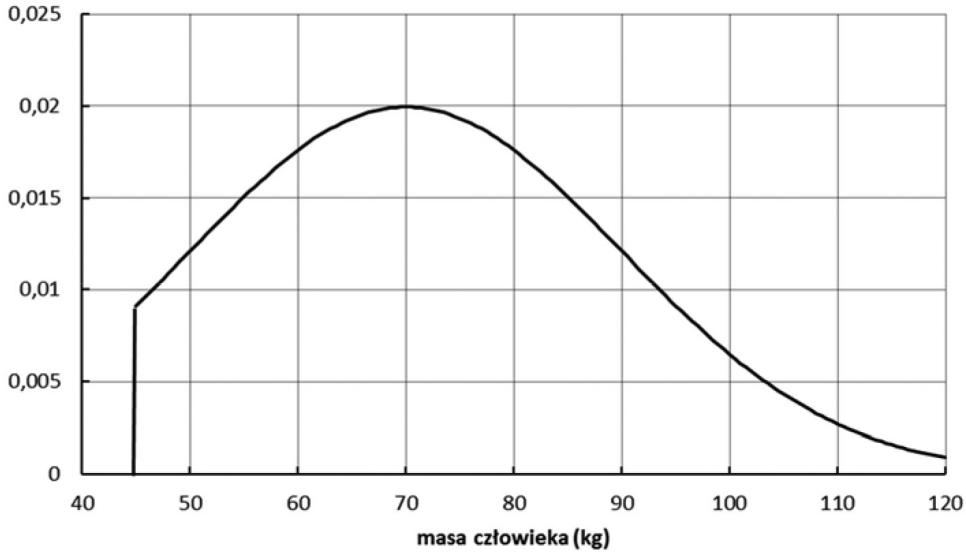
Rycina 5.3. Wykres funkcji gęstości rozkładu prawdopodobieństwa zmiennej: liczba dni pracy w roku

3. Rozkład normalny obcięty o parametrach: średnia 70,00, odchylenie standardowe 15,00, dobrany zakres od 30,00 do 95,00:



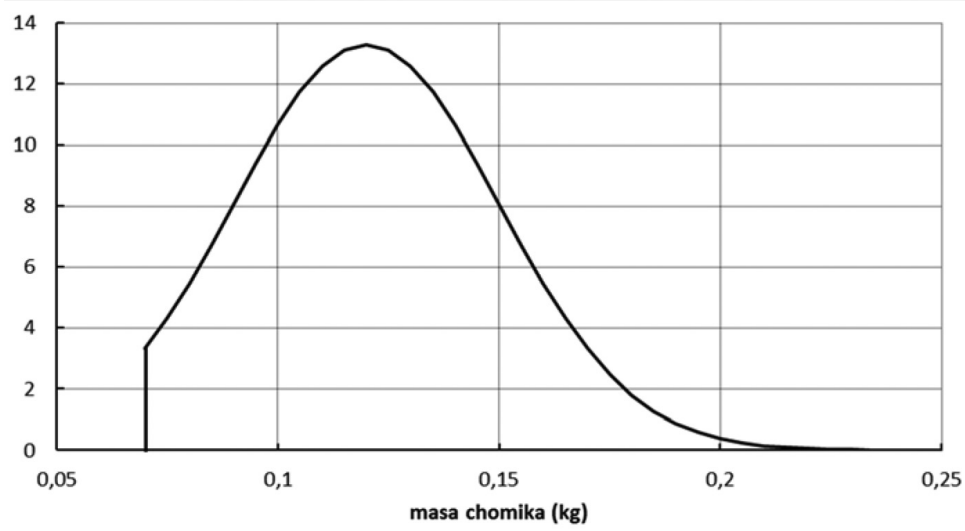
Rycina 5.4. Wykres funkcji gęstości rozkładu prawdopodobieństwa zmiennej: czas trwania życia człowieka

4. Rozkład normalny obcięty o parametrach: średnia 70,00, odchylenie standardowe 20,00, dobrany zakres od 45,00 do 120,00:



Rycina 5.5. Wykres funkcji gęstości rozkładu prawdopodobieństwa zmiennej: masa człowieka

5. Rozkład normalny obcięty o parametrach: średnia 0,12, odchylenie standardowe 0,03, dobrany zakres od 0,07 do 0,25:

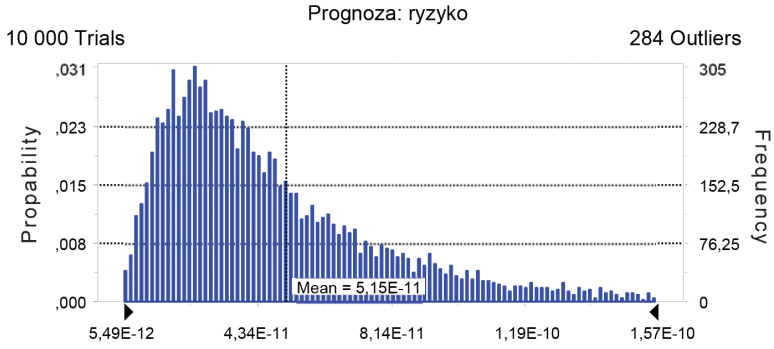


Rycina 5.6. Wykres funkcji gęstości rozkładu prawdopodobieństwa zmiennej: masa chomika

I teraz, stosując metodę Monte Carlo, zamiast jednej wartości ryzyka otrzymujemy rozkład prawdopodobieństwa ryzyk dla człowieka, ale ryzyk związanych z narażeniem na benzo(a)piren w stężeniu 0,0001 mg/m^3 przez okres 3 lat w warunkach narażenia zawodowego. Zwrot „prawdopodobieństwa ryzyk" brzmi może dziwacznie, gdyż jest to „prawdopodobieństwo prawdopodobieństwa", ale właśnie o to chodzi – nie otrzymujemy jednej liczby wyrażającej ryzyko, tylko cały zakres zmienności ryzyka, w tym przypadku ryzyka choroby nowotworowej będącą konsekwencją narażenia człowieka na benzo(a)piren. Charakterystyka tego rozkładu przedstawiona jest w tab. 5.5.

Tabela 5.5. Parametry rozkładu uzyskanego metodą Monte Carlo

Parametry statystyczne	Wartość
Liczebność próby	10 000
Średnia	5,15E-11
Mediana	4,00E-11
Moda	6,10E-11
Odchylenie standardowe	4,11E-11
Wariancja	1,69E-21
Współczynnik skośności	2,71E+00
Współczynnik spłaszczenia	1,66E+01
Współczynnik zmienności	7,97E-01
Minimum	2,94E-12
Maksimum	5,61E-10



Rycina 5.7. Charakterystyka ryzyka choroby nowotworowej będącej konsekwencją narażenia człowieka na benzo(a)piren w stężeniu 0,0001 mg/m^3 przez okres 3 lat w warunkach narażenia zawodowego

Tabela 5.6. Percentyle rozkładu ryzyka uzyskanego metodą Monte Carlo

Percentyl	Wartość ryzyka
0%	2,94E-12
10%	1,64E-11
20%	2,24E-11
30%	2,76E-11
40%	3,34E-11
50%	4,00E-11
60%	4,77E-11
70%	5,81E-11
80%	7,34E-11
90%	9,93E-11
100%	5,61E-10

Dzięki podejściu bayesowskiemu zamiast jednej wartości ryzyka otrzymujemy całe spektrum wraz z odpowiednimi prawdopodobieństwami ich wystąpienia. Jednak problemem jest tutaj konieczność określenia – bądź *a priori*, bądź przez oszacowanie – rozkładów prawdopodobieństwa używanych parametrów modelu. Zmiana tych rozkładów prowadzi do uzyskania, najczęściej całkowicie różnych, rezultatów końcowych. Zatem podejście bayesowskie również nie jest pozbawione wad.

5.3. Metody wiarygodnościowe (paradygmat wiarygodnościowy)

Podejście wiarygodnościowe jest podstawowym w modelowaniu przy użyciu regresji logistycznej. Przy oszacowaniu współczynników regresji w modelach liniowych wykorzystywana jest metoda najmniejszych kwadratów, natomiast w regresji logistycznej stosuje się metodę największej wiarygodności.

5.3.1. Zagadnienie estymacji

Metoda największej wiarygodności to jedna z metod estymacji, być może najważniejsza. Swą szczególną pozycję zawdzięcza cennym właściwościom, np. niezmienniczości, zgodności uzyskiwanych estymatorów (Pawłowski, 1976). Główna

idea metody największej wiarygodności polega na tym, aby za oceny szacowanych parametrów przyjmować takie wartości, przy których funkcja wiarygodności jest największa.

Niech x oznacza realizację pewnego ciągu obserwacji, a $f(x, \boldsymbol{\theta})$ oznacza funkcję gęstości; $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)$ jest parametrem wektorowym o wartościach w pewnym zbiorze Θ. Wiarygodność wektora $\boldsymbol{\theta}$ przy danej obserwacji x definiuje się jako funkcję:

$$L(\theta; x) \propto f(x; \theta) \tag{5.10}$$

Zgodnie z zasadą największej wiarygodności jako oszacowanie parametru $\boldsymbol{\theta}$ przyjmuje się taką wartość $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \cdots, \hat{\theta}_q)$ dla której:

$$L(\hat{\theta}|x) = \sup_{\theta \in \Theta} L(\theta|x) \tag{5.11}$$

Może się zdarzyć, że supremum nie jest realizowane – wtedy oszacowaniem „prawie największej wiarygodności" będzie $\boldsymbol{\theta}^*$ takie, że:

$$L(\theta^*|x) \geq c \cdot \sup_{\theta \in \Theta} L(\theta|x) \tag{5.12}$$

gdzie c jest ustaloną liczbą z przedziału $(0, 1)$.

W praktyce wygodniej jest posługiwać się funkcją $l(\theta|x) = \log L(\theta|x)$ i wówczas $\hat{\theta}$ określone wzorem (5.11) spełnia warunek:

$$l(\hat{\theta}|x) = \sup_{\theta \in \Theta} l(\theta|x) \tag{5.13}$$

W przypadku, gdy supremum we wzorze (5.13) realizowane jest przez pewien wewnętrzny punkt zbioru $\boldsymbol{\Theta}$ i funkcja $l(\theta|x)$ jest różniczkowalna względem $\boldsymbol{\theta}$, pochodne cząstkowe tej funkcji przyjmują wartość zero w punkcie supremum i jest rozwiązaniem układu równań:

$$\frac{\partial l(\theta|x)}{\partial \theta_i} = 0; \quad i = 1, 2, \cdots, q \tag{5.14}$$

Równania (5.14) nazywają się równaniami największej wiarygodności, a każde rozwiązanie tych równań jest oszacowaniem parametru $\boldsymbol{\theta}$ metodą największej wiarygodności. Funkcja $\hat{\theta}$ zdefiniowana wzorem (5.13), rozważana jako funkcja obserwacji x, nazywa się estymatorem największej wiarygodności (Rao, 1982).

Rozważmy następujący przykład (Gart i wsp., 1986).

Jakkolwiek by nie specyfikować parametrycznej postaci modelu dawka-odpowiedź, to prawdopodobieństwo zaobserwowania odpowiedzi przy narażeniu na substancję chemiczną w dawce d zależy od pewnych nieznanych parametrów. Ogólnie można przyjąć, że jest p takich parametrów $\theta_1, \ldots, \theta_p$ tworzących wektor

$\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$. Prawdopodobieństwo odpowiedzi będzie oznaczone jako $P^*(d; \boldsymbol{\theta})$. Dalej naszkicowany zostanie sposób szacowania nieznanych parametrów metodą największej wiarygodności na podstawie obserwowanych danych. Przyjmuje się, że w eksperymencie użyto n zwierząt podzielonych na $I + 1$ grup. Każda grupa narażona była na inny poziom badanej substancji chemicznej: $0 = d_0 < d_1 < \ldots < d_I$ oraz w każdej grupie n_i zwierząt u x_i rozwinął się nowotwór w okresie badania. Zakładając, że odpowiedź u każdego ze zwierząt pojawia się niezależnie od reakcji innych zwierząt biorących udział w eksperymencie, funkcja wiarygodności obserwowanych wyników w modelu dawka-odpowiedź $P^*(d; \boldsymbol{\theta})$ jest wyrażona wzorem:

$$L(\theta) = \prod_{i=o}^{l} \binom{n_i}{x_i} (P_i^*)^{x_i} (1 - P_i^*)^{n_i - x_i} \tag{5.15}$$

gdzie $P_i^* = P^*(d_i; \theta)$. Wartość $\hat{\theta}$ parametru $\boldsymbol{\theta}$, która maksymalizuje $L(\boldsymbol{\theta})$ jest nazywana oszacowaniem największej wiarygodności. Ponieważ maksymalizacja funkcji $L(\boldsymbol{\theta})$ jest na ogół niemożliwa przy użyciu bezpośrednich procedur analitycznych, estymator największej wiarygodności $\hat{\theta}$ parametru $\boldsymbol{\theta}$ zazwyczaj otrzymuje się przez zastosowanie numerycznych procedur iteracyjnych[1].

5.3.2. Zagadnienie testowania (Magiera, 2007; Lindgren, 1962)

Rozważmy problem testowania:

$$\begin{cases} H_0: \theta \in \Theta_0 \\ H_1: \theta \in \Theta_1 = \Theta \backslash \Theta_0 \end{cases} \tag{5.16}$$[2]

Niech $L(\theta; x)$ oznacza funkcję wiarygodności. Określamy funkcję:

$$\lambda^*(x) = \frac{\sup_{\theta \in \Theta_0} L(\theta; x)}{\sup_{\theta \in \Theta} L(\theta; x)} \tag{5.17}$$

gdzie x to zaobserwowana wartość zmiennej X. Licznik jest pomyślany jako najlepsze „wyjaśnienie" obserwowanego rezultatu w H_0, a mianownik jako najlepsze wyjaśnienie w H_1. Jeśli iloraz jest duży, to lepsze wyjaśnienie jest znajdowane w H_0 niż w H_1, natomiast jeśli iloraz jest mały, to lepsze wyjaśnienie jest znajdowane w H_1. Określenie wartości (pozwalającej zdecydować, jak duże jest „duże") znów jest problemem wyważenia ważności obu rodzajów błędów (pierwszego i drugiego rodzaju).

W wielu przypadkach wygodnie jest zmienić nieznacznie metodę. Zamiast λ^* można rozważać:

1 Fragment ten pochodzi z monografii Szymczaka (1999).

2 Symbol $\Theta \backslash \Theta_0$ oznacza odejmowanie zbiorów.

$$\lambda(x) = \frac{\sup_{\theta \in \Theta_0} L(\theta;x)}{\sup_{\theta \in \Theta} L(\theta;x)} \tag{5.18}$$

Jeśli $\lambda^* < 1$, wtedy $\lambda = \lambda^*$, ale jeśli $\lambda^* \geq 1$, wtedy $\lambda = 1$. Obszar odrzucenia zdefiniowany przez $\lambda < K$ jest taki sam, jak w przypadku $\lambda^* < K$, pod warunkiem, że $K \leq 1$. W praktyce zamiast na λ wygodniej jest pracować na $-\log \lambda$ i wówczas obszar krytyczny określany jest jako $-\log \lambda >$ pewnej stałej.

Gdy stosujemy test oparty na statystyce λ (albo na $-\log \lambda$), konieczna jest znajomość rozkładu prawdopodobieństwa λ. Jest to zazwyczaj bardzo skomplikowane i najczęściej wykorzystujemy informację o asymptotycznym rozkładzie statystyki $-2\log \lambda$. Jest to asymptotyczny rozkład chi-kwadrat.

W przypadku porównywania dwóch modeli logistycznych Hilbe (2009) definiuje statystykę będącą podstawą testu największej wiarygodności jako:

$$G = -2(LL_r - LL_f) \tag{5.19}$$

gdzie LL_f to logarytm funkcji wiarygodności modelu (modelu pełnego) rozszerzonego o pewne czynniki ryzyka w stosunku do modelu, dla którego logarytm funkcji wiarygodności jest oznaczony jako LL_r (modelu zredukowanego). Test ilorazu wiarygodności jest szczególnie przydatny w ocenie, czy czynnik ryzyka albo grupa czynników ryzyka w istotny sposób „bierze udział" w modelowaniu.

PRZYKŁAD 5.2 (Lindgren, 1962)

Rozważmy zagadnienie testowania:

$$\begin{cases} H_0: \mu = \mu_0 \\ H_1: \mu \neq \mu_o \end{cases} \tag{5.20}$$

na podstawie próby z populacji, w której badana cecha ma rozkład normalny o znanej wariancji σ^2. Funkcja wiarygodności ma wówczas postać:

$$L(\mu) = (2\pi\sigma^2)^{-n/2} exp\left[\frac{-1}{2\sigma^2}\sum(x_i - \mu)^2\right] \tag{5.21}$$

Maksimum po wszystkich μ jest osiągane dla $\hat{\mu} = \bar{x}$, oszacowanym metodą największej wiarygodności parametru μ, zatem mianownikiem w wyrażeniu dla λ jest $L(\bar{x})$. Licznik to $L(\mu_0)$, ponieważ H_0 składa się z pojedynczego punktu μ_0. Iloraz λ jest wówczas postaci $L(\bar{x})/L(\mu_0)$ oraz:

$$\log\ \lambda = -\log\ L(\mu_0) + \log\ L(\bar{x}) = \frac{1}{2\sigma^2}\{\sum(x_i - \mu_0)^2 - \sum(x_i - \bar{x})^2\} = \frac{n}{2\sigma^2}[\bar{x} - \mu_0]^2 \tag{5.22}$$

Obszar krytyczny –log λ < stałej jest zatem równoważny:

$$|\bar{x} - \mu_0| > \text{stałej} \tag{5.23}$$

Odpowiada on obszarowi krytycznemu statystyki chi-kwadrat z jednym stopniem swobody.

PRZYKŁAD 5.3

Zbudujmy dwa modele regresji logistycznej, w których zmienną wynikową z badania będzie dwustanowa zmienna „ukl_kraz" (0 – nie zdiagnozowano u pacjenta chorób układu krążenia, 1 – zdiagnozowano jedną z chorób układu krążenia) (dane: Dudek, 2007).

W pierwszym modelu (nazwę go modelem zredukowanym) jako czynniki ryzyka wystąpienia chorób układu krążenia znajdą się zmienne: średnie z dwóch pomiarów – ciśnienie skurczowe krwi (skur_sre), HDL, LDL i grupa wieku. Zmienna „grupa_wieku" jest zmienną dyskretną:

$$\text{grupa\_wieku} = \begin{cases} 1 & \text{do 35 lat} \\ 2 & 36 - 45 \text{ lat} \\ 3 & 46 \text{ albo więcej lat} \end{cases}$$

W tab. 5.7 przedstawione są wyniki modelowania.

Tabela 5.7. Zmienne w modelu regresji logistycznej

Zmienne		B	Błąd standardowy	Wald	*df*	Istotność	Exp(B)	95% przedział ufności dla EXP(B)	
								dolna granica	górna granica
Krok 1[a]	skur_sre	,088	,012	55,876	1	,000	1,092	1,067	1,117
	LDL	,005	,004	1,667	1	,197	1,005	,997	1,014
	grupa_wieku			6,354	2	,042			
	grupa_wieku(1)	,637	,276	5,313	1	,021	1,890	1,100	3,248
	grupa_wieku(2)	,755	,393	3,701	1	,054	2,128	,986	4,594
	HDL	–,007	,010	,470	1	,493	,993	,974	1,013
	Stała	–12,775	1,786	51,183	1	,000	,000		

[a] Zmienne wprowadzone w kroku 1: skur_sre, LDL, grupa_wieku, HDL.

Tabela 5.8. Wartość funkcji wiarygodności i współczynniki pseudo-R^2 dla modelu pokazanego w tab. 5.7

Krok	–2 logarytm wiarygodności	*R*-kwadrat Coxa i Snella	*R*-kwadrat Nagelkerke'a
1	403,374[a]	,226	,319

[a] Estymacja została zakończona na iteracji o numerze 5, ponieważ oszacowania parametrów zmieniły się o mniej niż ,001.

Do otrzymanego modelu włączono zmienną „rozk_sre" (średnią z dwóch pomiarów ciśnienia rozkurczowego). Model ten nazwałem modelem pełnym. Uzyskano następujące rezultaty.

Tabela 5.9. Zmienne w rozszerzonym o zmienną „rozk_sre" modelu regresji logistycznej

Zmienne		B	Błąd standardowy	Wald	*df*	Istotność	Exp(B)	95% przedział ufności dla EXP(B)	
								dolna granica	górna granica
Krok 1	HDL	–,011	,011	,960	1	,327	,989	,969	1,011
	LDL	,005	,004	1,129	1	,288	1,005	,996	1,013
	skur_sre	–,008	,021	,136	1	,712	,992	,952	1,035
	rozk_sre	,168	,033	25,804	1	,000	1,183	1,109	1,262
	grupa_wieku			6,802	2	,033			
	grupa_wieku(1)	,648	,290	4,984	1	,026	1,911	1,082	3,375
	grupa_wieku(2)	,916	,420	4,747	1	,029	2,499	1,096	5,696
	stała	–13,878	1,843	56,709	1	,000	,000		

Tabela 5.10. Wartość funkcji wiarygodności i współczynniki pseudo-R^2 dla rozszerzonego modelu pokazanego w tab. 5.9

Krok	–2 logarytm wiarygodności	*R*-kwadrat Coxa i Snella	*R*-kwadrat Nagelkerke'a
1	371,500	,283	,400

Logarytm wiarygodności (*log likelihood*), czyli –2 $LL_{zredukowany}$ wynosi 403,374, zaś –2 $LL_{pełny}$ równa się 371,500.

$$-2(LL_{zredukowany} - LL_{pełny}) \tag{5.24}$$

Wyrażenie opisane wzorem (5.24) ma rozkład chi-kwadrat z liczbą stopni swobody będącą różnicą liczby czynników ryzyka w obu modelach (Hilbe, 2009; Hosmer, Lemeshow, 1989). Wartość statystyki chi-kwadrat to:

403,374 – 371,500 = 31,874
liczba stopni swobody: 1

Prawdopodobieństwo odpowiadające tej wartości statystyki jest równe $0{,}165 \cdot 10^{-7} < 0{,}05$. Oznacza to, że wprowadzona do modelu zmienna „rozk_sre" jest zmienną w istotny sposób związaną z prawdopodobieństwem wystąpienia chorób układu krążenia. Wartości *LL* służą w tym przykładzie do porównania obu modeli i ewentualnie podjęcia decyzji, czy zmienną „rozk_sre" można usunąć z modelu pełnego.

Jak zauważa Hilbe (2009), wyniki testu Walda i testu największej wiarygodności dla pojedynczego czynnika ryzyka nie zawsze są spójne.

Podsumowanie

Jak dało się zauważyć przy omawianiu metody bayesowskiej, jej stosowanie wymusza określenie *a priori* rozkładów prawdopodobieństwa pewnych parametrów. I jest to główny zarzut podnoszony przez statystyków niebayesowskich. Jednak Royall (2000a, 2000b), omawiając kłopoty związane ze stosowaniem statystyki bayesowskiej, formułuje dwa ważne pytania odnoszące się do wszelkich działań statystycznych. Mianowicie, używając metodologii Neymana-Pearsona, próbujemy odpowiedzieć na pytanie: „Co powinienem zrobić?" zamiast: „Co »mówią« te dane?", a w odniesieniu do metod bayesowskich: „Co powinienem sądzić?" albo „W co powinienem wierzyć?".

W podobnym duchu wypowiada się Lindley (1961) (zauważmy, od jak dawna toczy się ta dyskusja):

> Większość aktualnych pomysłów statystycznych nie używa takiego rozkładu [rozkładu prawdopodobieństwa *a priori* – W.Sz.], lecz czujemy, że pewna iluminacja oraz zrozumienie mogą być korzystne poprzez ich wykorzystanie, i w pełnych uniesienia momentach nawet poczujemy, że tylko w pełni bayesowska postawa w myśleniu statystycznym jest spójna i praktyczna[1].

W swojej ogólności sformułowania te brzmią nawet ciekawie i obiecująco, lecz sądzę, iż obiecują zdecydowanie zbyt dużo. Pierwsze postawione powyżej pytanie: „Co powinienem zrobić?", o tyle wydaje się niewłaściwe, że najpierw powinniśmy zapytać: „Co chcę zrobić?", a następnie zastanawiać się, jak zrobić to najlepiej. Oczywiście trzeba wykorzystać w tym celu całą dostępną wiedzę, a więc również informacje o postaci ewentualnych rozkładów prawdopodobieństwa, jeśli takie informacje istnieją i mają dla nas sens. Pytanie: „Co mówią te dane?" jest niczym nie uzasadnioną antropomorfizacją danych – decyzje podejmuje zawsze prowadzący badania, interpretujący otrzymywane dane i zależności między nimi.

---

1 „Most current statistical thinking does not use such a distribution (prior probability distribution – przyp. W.Sz.) but we feel that some illumination and understanding can be gained through its use, and, in our more enthusiastic moments, we even feel that only a completely Bayesian attitude towards statistical thinking is coherent and practical".

„W co powinienem wierzyć?" – trudno zgodzić się z tak sformułowanym pytaniem. Możemy stawiać pewne hipotezy, zgadywać, że jest tak, a nie inaczej, co sugeruje Lindley (1961), ale zawsze mamy obowiązek zweryfikować nasze przypuszczenia i nic nas z tego obowiązku nie zwalnia, inaczej przestanie być to nauka.

Czy na podstawie otrzymanych wyników analizy statystycznej uzyskujemy jakąś prawdziwą wiedzę? Jednoznacznej odpowiedzi udzielają Berger i Berry (1988) w publikacji pod symptomatycznym tytułem *Analiza statystyczna i iluzja obiektywności*: „Pogodzenie się z subiektywnością analizy statystycznej byłoby zdrowe dla nauki jako całości […]"[2].

Praktyczne stosowanie jakiejkolwiek z metod statystycznych jest wnioskowaniem indukcyjnym, w którym uogólniamy przypadek szczegółowy. To proces podejmowania decyzji w warunkach niepełnej informacji, zatem podjęta decyzja jest obarczona jakąś niepewnością. Uzyskanie takiej niepewnej wiedzy i ocena rozmiaru niepewności prowadzi do zdobycia wiedzy użytecznej, choć nie jest to wiedza pewna (Rao, 1994).

Jeszcze raz odwołam się do książki Bromek i Pleszczyńskiej (1988): „[…] na ogół nie potrafimy rozwiązywać problemów statystycznych, o które chodzi nam naprawdę; robimy zatem unik, obierając fikcyjne modele i formułując sztuczne problemy zastępcze, które rozwiązać umiemy. Dzisiejsza metodyka statystyczna poprzestaje na wyrażaniu nadziei, że rozwiązania te są dostatecznie dobre w danej sytuacji praktycznej". Mimo że od napisania tych słów minęło już 30 lat, to niestety wydają się nadal aktualne.

I jeszcze raz Rao (1994): „Statystyka jest bardziej sposobem myślenia lub wnioskowania niż pęczkiem recept na młócenie danych w celu odsłonięcia odpowiedzi".

Zatem trzeba myśleć, myśleć, myśleć o tym, co chce się zrobić i niezbędna wydaje się bardzo ścisła współpraca badacza merytorycznego i statystyka, choć – szczególnie w początkowym okresie – dla obu stron jest ona bardzo, bardzo trudna.

2 „Acknowledging the subjectivity of statistical analysis would be healthy for science as a whole […]".

Bibliografia

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

Ahad N. A., Yin T. S., Othman A. R., Yaacob C. R. (2011): *Sensitivity of Normality Tests to Non-normal Data*. „Sains Malaysiana", 40, 6, s. 637–641.

Aitkin M. (2011): *Paradigms for Statistical Inference*, https://researchers.ms.unimelb.edu.au/~maitkin@unimelb/paradigms2.pdf, https://pdfs.semanticscholar.org/5226/77da13ad4ee11b2e8a9a02b0c1bdf6a77f4f.pdf (dostęp: 29.08.2011).

Allen J., Le H. (2007): *An Additive Measure of Overall Effect Size for Logistic Regression Models*. „Journal of Educational and Behavioral Statistics", http://jebs.aera.net (dostęp: 1.12.2008).

Anscombe F. J., Aumann R. J. (1963): *A Definition of Subjective Probability.* „The Annals of Mathematical Statistics", 34, 1, s. 199–205.

APA (2010): *Publication Manual. Sixth Edition.* American Psychological Association, Washington.

Armitage P., Doll R. (1954): *The Age Distribution of Cancer and a Multi-stage Theory of Carcinogenesis.* „British Journal of Cancer", 8, 1, s. 1–12.

Armitage P., Doll R. (1961): *Stochastic Models for Carcinogenesis*. „Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, University of California Press", 4, s. 19–38.

Armstrong J. S. (2007): *Significance Tests Harm Progress in Forecasting*. „International Journal of Forecasting", 23, 2, s. 321–327.

Awan H. M. (2001): *Effect of Departure from Standard Assumptions Used in Analysis of Variance.* „Journal of Research (Science)", 12, 2, s. 180–188.

Babu G. J., Padmanabhan A. R., Puri M. L. (1999): *Robust One-Way ANOVA under Possibly Non-Regular Conditions*. „Biometrical Journal", 41, 3, s. 321–339.

Bakeman R. (2005): *Recommended Effect Size Statistics for Repeated Measures Designs*. „Behavior Research Methods", 37, 3, s. 379–384.

Bandyopadhyay P. S., Forster M. R. (red.) (2011): *Handbook of Philosophy of Science*, Vol. 7: *Philosophy of Statistics*. Elsevier, Amsterdam.

Bernardo J. M. (2011): *Modern Bayesian inference: foundations and objective methods*. W: *Handbook of Philosophy of Science*, Vol. 7: *Philosophy of Statistics.* Red. P. S. Bandyopadhyay, M. R. Forster, s. 263–306. Elsevier, Amsterdam.

Berger J. O. (2003): *Could Fisher, Jefreys and Neyman Have Agreed on Testing?* „Statistical Sciences", 18, 1, s. 1–32.

Berger J. O., Berry D. A. (1988): *Statistical Analysis and the Illusion of Objectivity.* „American Scientist", 76, s. 159–165.

Blalock H. M. (1975): *Statystyka dla socjologów*. Państwowe Wydawnictwo Naukowe, Warszawa.

Blume J. D. (2002): *Likelihood Methods for Measuring Statistical Evidence.* „Statistics in Medicine", 21, s. 2563–2599.

Bock R. D. (1975): *Multivariate Statistical Methods in Behavioral Research*. McGraw-Hill Inc., New York.

Bromek T., Pleszczyńska E. (red.) (1988): *Teoria i praktyka wnioskowania statystycznego*. Państwowe Wydawnictwo Naukowe, Warszawa.

Brown M. B., Forsythe A. B. (1974): *Robust Tests for the Equality of Variances*. „Journal of the American Statistical Association", 69, s. 364–367.

Calculating Cancer Risk Due to Occupational Exposure to Genotoxic Carcinogens (1995). Report of the Dutch Expert Committee on Occupational Standards, a Committee of the Health Council of the Netherlands, No. 1995/06WGD, The Hague, 18 October 1995.

Carlin B. P., Louis T. A. (2000): *Bayes and Empirical Bayes Methods for Data Analysis.* Chapman and Hall/CRC, Boca Raton.

Carver R. P. (1993). *The Case against Statistical Significance Testing, Revisited*. „Journal of Experimental Education", 1993, 61, s. 287–292. Cyt. za: A. J. Onwuegbuzie, N. L. Leech (2004): *Post Hoc Power: A Concept Whose Time Has Come.* „Understanding Statistics", 3, 4, s. 201–230.

Chinn S. (2000): *A Simple Method for Converting an Odds Ratio to Effect Size for Use in Meta-analysis.* „Statistics in Medicine", 19, 22, s. 3127–3131.

Chmura-Kraemer H., Kupfer D. J. (2006): *Size of Treatment Effects and Their Importance to Clinical Research and Practice.* „Biological Psychiatry", 56, 11, s. 90–996.

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

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

Cohen J. (1992a): *A Power Primer*. „Psychological Bulletin", 112, 1, s. 155–159.

Cohen J. (1992b): *Statistical Power Analysis*. „Current Directions in Psychological Sciences", 1, 3, s. 98–101.

Cohen J. (1994): *The Earth Is Round ($p < .05$)*. „American Psychologist", 49, 12, s. 997–1003.

Collins J. F., Brown J. P., Dawson S. V., Marty M. A. (1991): *Risk Assessment for Benzo*[*a*]*pyrene*. „Regulatory Toxicology and Pharmacology", 13, 2, s. 170–184.

Congdon P. (2002): *Bayesian Statistical Modelling.* John Wiley and Sons, Chichester.

Connett J. E., Smith Judith A., McHugh R. B. (1987): *Sample Size and Power for Pair-matched Case-control Studies.* „Statistics in Medicine", 6, s. 53–59.

Connor R. J. (1987): *Sample Size for Testing Differences in Proportions for the Paired-sample Design.* „Biometrics", 43, s. 207–211.

Conover W. J., Johnson M. E., Johnson M. M. (1981): *A Comparative Study of Tests for Homogeneity of Variances, with Applications to the Outer Continental Shelf Bidding Data*. „Technometrics", 23, s. 351–361.

Cumming G. (2014): *The New Statistics: Why and How.* „Psychological Science ", 25, 1, s. 7–29.

De Kleijn W. P. E., De Vries J., Petal A. H. M., Wijnen P. A. H. M., Drent M. (2011): *Minimal (clinically) important differences for the Fatigue Assessment Scale in sarcoidosis*, „Respiratory Medicine", 105, 1, s. 1388–1395.

Denis D. J. (2003): *Alternatives to Null Hypothesis Significance Testing*. "Theory and Science", 4, 1, s. 1–17, http://theoryandscience.icaap.org/content/vol14.1/02_denis.html (dostęp: 6.12.2007).

Dienes Z. (2011): *Bayesian Versus Orthodox Statistics: Which Side Are You On?* „Perspective on Psychological Science", 6, 3, s. 274–290.

Donoghue D., Physiotherapy Research and Older People (PROP), Stokes E. K. (2009): *How Much Change is True Change? The Minimum Detectable Change of the Berg Balance Scale in Elderly People.* „Journal of Rehabilitation Medicine", 41, s. 343–346.

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

Efron B. (1978): *Controversies in the Foundation of Statistics.* „American Mathematical Monthly", 85, 4, s. 231–246.

Efron B. (1998): *R. A. Fisher in the 21st Century. Invited Paper Presented at the 1996 R. A. Fisher Lecture.* „Statistical Science", 13, 2, s. 95–122.

Eisen S. V., Ranganathan G., Seal P., Spiro A. III (2007): *Measuring Clinically Meaningful Change Following Mental Health Treatment.* „The Journal of Behavioral Health Services and Research", 34, 3, s. 272–289.

Field A. (2009): *Discovering Statistics Using SPSS. Third Edition*. Sage Publications, Thousand Oaks.

Fisher R. A. (1922): *On the Mathematical Foundations of Theoretical Statistics.* „Philosophical Transactions of the Royal Society of London". Series A, Containing Papers of a Mathematical or Physical Character, 222, s. 309–368.

Fisher R. A. (1935): *The Logic of Inductive Inference (with Discussion)*. „Journal of the Royal Statistical Society", 98, 1, s. 39–82.

Fisher R. A. (1955): *Statistical Methods and Scientific Induction.* „Journal of the Royal Statistical Society". Series B (Methodological), 17, 1, s. 69–78.

Fisher R. A. (1956): *Statistical Methods and Scientific Inference.* Oliver and Bond, Edinburgh. Cyt. za: G. Gigerenzer (2004): *Mindless Statistics.* „Journal of Behavioral and Experimental Economics" (formerly „The Journal of Socio-Economics"), 33, 5, s. 587–606.

Fisz M. (1969): *Rachunek prawdopodobieństwa i statystyka matematyczna*. Państwowe Wydawnictwo Naukowe, Warszawa.

Frans F. A., Nieuwkerk P. T., Met R., Bipat S., Legemate D. A., Reekers J. A., Koelemay M. J. W. (2014): *Statistical or Clinical Improvement? Determining the Minimally Important Difference for the Vascular Quality of Life Questionnaire in Patients with Critical Limb Ischemia.* „European Journal of Vascular and Endovascular Surgery", 47, 2, s. 180–186.

Freedman D. (1995/1996): *Some Issues in the Foundations of Statistics.* „Foundations of Science", 1, s. 19–39.

Freeman G. H., Halton J. H. (1951): *Note on an Exact Treatment of Contingency, Goodness of Fit and Other Problems of Significance.* „Biometrika", 38, s. 141–149.

Games P. A. (1971): *Multiple Comparisons of Means.* „American Educational Research Journal", 8, s. 531–565.

Garcia-Perez A. (2008): *Approximations for F-Tests Which Are Ratios of Sums of Squares of Independent Variables with a Model Close to Normal.* „Test", 17, 2, s. 350–369.

Garson G. D. (2012): *Testing Statistical Assumptions.* Statistical Associates Publishing, North Carolina State University, Raleigh.

Gart J. J., Krewski D., Lee P. N., Tarone R. E., Wahrendorf J. (1986): *Statistical Methods in Cancer Research*, Vol. III: *The Design and Analysis of Long-term Animal Experiments.* International Agency for Research on Cancer, Lyon.

Gastwirth J. L., Gel Yulia R., Miao W. (2009): *The Impact of Levene's Test of Equality of Variances on Statistical Theory and Practice.* „Statistical Science", 24, 3, s. 343–360.

Gigerenzer G. (2004): *Mindless Statistics.* „Journal of Behavioral and Experimental Economics" (formerly „The Journal of Socio-Economics"), 33, 5, s. 587–606.

Gigerenzer G., Marewski J. N. (2015): *Surrogate Science: The Idol of a Universal Method for Scientific Inference.* „Journal of Management", 41, 2, s. 421–440.

Gigerenzer G., Krauss S., Vitouch O. (2004): *The Null Ritual. What You Always Wanted to Know About Significance Testing but Were Afraid to Ask.* W: *The Sage Handbook of Quantitative Methodology for the Social Sciences.* Red. D. Kaplan, s. 391–408. Sage Publications, Thousand Oaks.

Gillett R. (1994): *Post Hoc Power Analysis.* „Journal of Applied Psychology", 79, 5, s. 783–785.

Glass G. V., Peckham P. D., Sanders J. R. (1972): *Consequences of Failure to Meet Assumptions Underlying the Fixed Effects Analyses of Variance and Covariance.* „Review of Educational Research", 42, 3, s. 237–288.

Gomez J. (2012): *Tam gdzie proste są krzywe. Geometrie nieeuklidesowe.* RBA Coleccionables, Barcelona.

Gray C. D., Kinnear P. R. (2012): *IBM SPSS Statistics 19 Made Simple*. Psychology Press, New York. Cyt. za: K. L. Wuensch (2015): *Nonparametric Effect Size Estimators*. Materiały dydaktyczne. East Carolina University Department of Psychology, core.ecu.edu/psyc/wuenschk/docs30/Nonparametric-EffectSize.pdf (dostęp: 16.06.2015).

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, s. 203–208.

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, s. 387–392.

Greń J. (1968): *Modele i zadania statystyki matematycznej.* Państwowe Wydawnictwo Naukowe, Warszawa.

Hagen R. L. (1997): *In Praise of the Null Hypothesis Statistical Test*. „American Psychologist", 52, 1, s. 15–24.

Häggström O. (2012): *Why the Empirical Sciences Need Statistics So Desperately*. W: *European Congress of Mathematics, Krakow, 2–7 July, 2012.* Red. R. Latała, A. Ruciński,

P. Strzelecki, J. Świątkowski, D. Wrzosek, P. Zakrzewski, s. 347–360. European Mathematical Society Publishing House, Zurich.

Hartley, H. O. (1950): *The Maximum F-rates as a Shortcut Test for Heterogeneity of Variance.* „Biometrika", 37, 308–312.

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

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, s. 19–24.

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

Hsieh F. Y. (1989): *Sample Size Tables for Logistic Regression.* „Statistics in Medicine", 8, s. 795–802.

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, s. 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, s. 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, s. 2–11.

Jacobson N. S., Truax P. (1991): *Clinical Significance: A Statistical Approach to Defining Meaningful Change in Psychotherapy Research.* „Journal of Consulting and Clinical Psychology", 59, 1, s. 12–19.

Jeffreys H. (1961): *Theory of Probability*, 3rd ed. Oxford University Press, Oxford. Cyt. za: J. O. Berger (2003): *Could Fisher, Jefreys and Neyman Have Agreed on Testing?* „Statistical Sciences", 18, 1, s. 1–32.

Jones L. V., Tukey J. W. (2000): *A Sensible Formulation of the Significance Test.* „Psychological Methods", 5, 4, s. 411–414.

Karni E. (1993): *A Definition of Subjective Probabilities with State-dependent Preferences.* „Econometrica", 61, 1, s. 187–198.

Kerby D. S. (2014): *The Simple Difference Formula: An Approach to Teaching Nonparametric Correlation.* „Innovative Teaching", 3, 1. Cyt. za: K. L. Wuensch (2015): *Nonparametric Effect Size Estimators*. Materiały dydaktyczne. East Carolina University, Department of Psychology, Greenville.

Keselman H. J., Cribbie R. A., Wilcox R. R. (2002): *Pairwise Multiple Comparison Tests When Data Are Nonnormal.* „Educational and Psychological Measurement", 62, 3, s. 420–434.

Khan A., Rayner G. D. (2003): *Robustness to Non-Normality of Common Tests for the Many-Sample Location Problem.* „Journal of Applied Mathematics and Decision Sciences", 7, 4, s. 187–206.

Killeen P. R. (2005): *An Alternative to Null-Hypothesis Significance Tests*. „Psychological Science", 16, 5, s. 345–353.

Kirk R. E. (1994): *Experimental Design: Procedures for the Behavioral Sciences*, 3rd ed. Wadsworth Publishing, Belmont CA.

Kochanski G. (2005): *Brute Force as a Statistical Tool*, http://kochanski.org/gpk (dostęp: 1.03.2005).

Kołmogorow A. N. (1933): *Grundbegriffe der Wahrscheinlichkeitsrechnung*. „Ergebnisse der Mathematik und ihrer Grenzgebiete", 3, s. 1–62. Cyt. za: H. Bauer (1968): *Probability Theory and Elements of Measure Theory*. Holt, Rinehart and Winston Inc., New York.

Krämer W. (2011): *The Cult of Statistical Significance. What Economists Should and Should Not Do to Make Their Data Talk.* Working Paper Series of the German Data Forum (RatSWD). Working Paper No. 176, April 2011.

Kubot A. (2017): *Zastosowanie radialnej fali uderzeniowej i ultradźwięków w leczeniu zespołu łokcia tenisisty*. Rozprawa na stopień doktora nauk medycznych, Uniwersytet Medyczny, Łódź.

Kubot A., Grzegorzewski A., Synder M., Szymczak W., Kozłowski P. (2017): *Zastosowanie radialnej fali uderzeniowej i ultradźwięków w leczeniu zespołu łokcia tenisisty*. „Ortopedia Traumatologia Rehabilitacja", 19, 5(6), s. 415–426.

Kuhn T. S. (2009): *Struktura rewolucji naukowych.* Wydawnictwo Aletheia, Warszawa. Tłum. z: *The Structure of Scientific Revolutions*, 3rd ed. The University of Chicago Press, Chicago–London 1996.

Kyburg H. E., Jr. (1977): *Decisions, Conclusions, and Utilities.* „Synthese", 36, s. 87–96.

Lachin J. M. (1986): *Evaluation of Sample Size and Power for Analyses of Survival with Allowance for Nonuniform Patient Entry, Losses to Follow-up, Noncompliance, and Stratification.* „Biometrics", 42, s. 507–519.

Lambdin C. (2012): *Significance Tests as Sorcery: Science Is Empirical – Significance Tests Are Not.* „Theory and Psychology", 22, 1, s. 67–90.

Lantz B. (2013): *The Impact of Sample Non-normality on ANOVA and Alternative Methods.* „British Journal of Mathematical and Statistical Psychology", 66, s. 224–244.

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

Lee H. B., Katz G. S., Restori A. F. (2010): *A Monte Carlo Study of Seven Homogeneity of Variance Tests.* „Journal of Mathematics and Statistics", 6, 3, s. 359–366.

Lee M. D., Wagenmakers E.-J. (2005): *Bayesian Statistical Inference in Psychology: Comment on Trafimow (2003).* „Psychological Review", 112, 3, s. 662–668.

Lehmann E. L. (1959): *Testing Statistical Statistics.* John Wiley and Sons, New York. Pol. tłum.: *Testowanie hipotez statystycznych.* Państwowe Wydawnictwo Naukowe, Warszawa 1968.

Lehmann E. L. (1968): *Testowanie hipotez statystycznych.* Państwowe Wydawnictwo Naukowe, Warszawa.

Lehmann E. L. (1993): *The Fisher, Neyman-Pearson Theories of Testing Hypotheses: One Theory or Two?* „Journal of the American Statistical Association", 88, 424, s. 1242–1249.

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

Lehmann E. L. (2006): *Nonparametrics. Statistical Methods Based on Ranks.* Springer, New York.

Lemeshow S., Hosmer D. W., Klar J. (1988): *Sample Size Requirements for Studies Estimating Odds Ratios or Relative Risks.* „Statistics in Medicine", 7, s. 759–764.

Lenth R. V. (2007): *Post Hoc Power: Tables and Commentary.* Technical Report No. 378, The University of Iowa, Department of Statistics and Actuarial Sciences, s. 1–13.

Levene H. (1960): *Robust Tests of Equality of Variances*. W: *Contributions to Probability and Statistics. Essays in Honor of Harold Hotelling*, Red. I. Olkin, S. G. Ghurye, W. Hoeffding, W. G. Madow, H. B. Mann, s. 278–292. Stanford University Press, Palo Alto.

Leventhal L., Huynh C. L. (1996): *Directional Decisions for Two-Tailed Tests: Power, Error Rates and Sample Size.* „Psychological Methods", 1, 3, s. 278–292.

Levine T. R., Hullett C. R. (2002): *Eta Squared, Partial Eta Squared, and Misreporting of Effect Size in Communication Research.* „Human Communication Research", 28, 4, s. 612–625.

Levine T. R., Weber R., Hullet C., Park H. S., Lindsey Lisa L. M. (2008): *A Critical Assessment of Null Hypothesis Significance Testing in Quantitative Communication Research.* "Human Communication Research", 34, 171–187.

Levy P. S., Lemeshow S. (1991): *Sampling of Populations: Methods and Applications.* John Wiley and Sons, New York.

Lillestøl J. (2014): *Statistical Inference: Paradigms and Controversies in Historic Perspective.* NHH Norwegian School of Economics, Bergen, Department of Business and Management Science, https://www.nhh.no/globalassets/departments/business-and-management-science/statistical_inference.pdf (dostęp: 16.06.2015).

Lim T.-S., Loh W.-Y. (1996): *A Comparison of Tests of Equality of Variances.* „Computational Statistics and Data Analysis", 22, s. 287–301.

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

Lindley D. V. (1961): *The Use of Prior Probability Distributions in Statistical Inference and Decisions.* Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, Berkeley–Los Angeles.

Lindquist E. F. (1938): *A First Course in Statistics*. Houghton Mifflin Harcourt**,** Cambridge. Cyt za: Huberty C. J. (1993): *Historical Origins of Statistical Testing Practices: The Treatment of Fisher Versus Neyman-Pearson Views in Textbooks*. „Journal of Experimental Education", 61, 4, s. 317–333.

Loftus G. R. (1996): *Psychology Will Be a Much Better Science When We Change the Way We Analyze Data.* „Current Directions in Psychological Science", 5, 6, s. 161–171.

Machina M. J., Schmeidler D. (1992): *A More Robust Definition of Subjective Probability.* „Econometrica", 60, 4, s. 745–780.

Magee L. (1990): *R^2 Measures Based on Wald and Likelihood Ratio Joint Significance Tests.* „The American Statistician", 44, 3, s. 250–253.

Magiera R. (2007): *Modele i metody statystyki matematycznej.* Część II: *Wnioskowanie statystyczne*, wydanie drugie rozszerzone. Oficyna Wydawnicza GiS, Wrocław.

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

Mara C. A. (2013): *Testing for Equivalence of Group Variances. A Dissertation Submitted to the Faculty of Graduate Studies in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy.* York University, Toronto.

Maslow A. H. (1966): *The Psychology of Science.* Harper and Row, New York.

Mehlman M. H. (2017): ANOVA: *Analysis of Variance.* University of New Haven, http://math.newhaven.edu/mhm/courses/bstat/items.html, http://math.newhaven.edu/mhm/courses/estat/slides/ANOVA.pdf (dostęp: 4.12.2017).

Mehta C. R., Patel N. R. (1983). *A Network Algorithm for Performing Fisher's Exact Test in $r \times c$ Contingency Tables*. „Journal of the American Statistical Association", 78, 382, s. 427–434.

Mehta C. R., Patel N. R. (1986): *Algorithm 643. FEXACT: A Fortran Subroutine for Fisher's Exact Test on Unordered r × c Contingency Tables.* „ACM Transactions on Mathematical Software", 12, s. 154–161.

Menard S. (2000): *Coefficients of Determination for Multiple Logistic Regression Analysis.* „The American Statistician", 54, 1, s. 17–24.

Miller W. R., Manuel J. K. (2008): *How Large Must a Treatment Effect Be Before it Matters to Practitioners? An Estimation Method and Demonstration.* „Drug and Alcohol Review", 27, s. 524–528.

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

Moolgavkar S. H. (1986): *Carcinogenic Modeling: From Molecular Biology to Epidemiology*. „The Annual Review of Public Health", 7, s. 151–169.

Nagelkerke N. J. D. (1991): *A Note on a General Definition of the Coefficient of Determination*. „Biometrika", 78, 3, s. 691–692.

Neveu J. (1964): *Bases mathématiques du calcul des probabilités.* Masson et Cie, Paris.

Neyman J. (1969): *Zasady rachunku prawdopodobieństwa i statystyki matematycznej.* Państwowe Wydawnictwo Naukowe, Warszawa.

Neyman J. (1977): *Frequentist Probability and Frequentist Statistics.* „Synthese", 36, s. 97–131.

Neyman J., Pearson E. S. (1933): *On the Problem of the Most Efficient Tests of Statistical Hypotheses.* „Philosophical Transactions of the Royal Statistical Society. Section A", 231, s. 289–337. Cyt. za: E. L. Lehmann (1995): *Neyman's Statistical Philosophy.* „Probability and Mathematical Statistics", 15, s. 29–36.

Nordstokke D. W., Zumbo B. D. (2007): *A Cautionary Tale About Levene's Tests for Equal Variances*. „Journal of Educational Research and Policy Studies", 7, 1, s. 1–14.

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.* „Communication Methods and Measures", 1, 4, s. 291–299.

Onwuegbuzie A. J., Leech N. L. (2004): *Post Hoc Power: A Concept Whose Time Has Come.* „Understanding Statistics", 3, 4, s. 201–230.

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

Parra-Frutos I. (2009): *The Behaviour of the Modified Levene's Test When Data Are Not Normally Distributed.* „Computational Statistics", 24, s. 671–693.

Pawłowski Z. (1976): *Statystyka matematyczna*. Państwowe Wydawnictwo Naukowe, Warszawa.

Portier C. J., Kopp-Schneider A., Sherman C. D. (1993): *Using Cell Replication Data in Mathematical Modeling in Carcinogenesis.* „Environ Health Perspect" 101 (Suppl. 5), s. 79–86.

Rao C. R. (1965): *Linear Statistical Inference and Its Applications.* John Wiley and Sons, New York. Pol. tłum.: *Modele liniowe statystyki matematycznej*. Państwowe Wydawnictwo Naukowe, Warszawa, 1982.

Rao C. R. (1994): *Statystyka i prawda*. Wydawnictwo Naukowe PWN, Warszawa.

Rasch D. (2012): *Hypothesis Testing and the Error of the Third Kind.* „Psychological Test and Assessment Modeling", 54, 1, s. 90–99.

Razali N. M., Wah Y. B. (2011): *Power Comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling Tests.* „Journal of Statistical Modeling and Analytics", 2, 1, s. 21–33.

Richardson J. T. E. (1996): *Measures of Effect Size.* „Behavior Research Methods, Instruments and Computers", 28, 1, s. 12–22.

Richardson J. T. E. (2011): *Eta Squared and Partial Eta Squared as Measures of Effect Size in Educational Research.* „Educational Research Review", 6, s. 135–147.

Roberts S., Pashler H. (2000): *How Persuasive Is a Good Fit? A Comment on Theory Testing.* „Psychological Review", 107, 2, s. 358–367.

Robinson D. H., Wainer H. (2001): *On the Past and Future of Null Hypothesis Significance Testing*. Research Report RR-01-24, Educational Testing Service, Princeton.

Rodgers J. L. (2010): *The Epistemology of Mathematical and Statistical Modeling. A Quiet Methodological Revolution.* „American Psychologist", 65, 1, s. 1–12.

Rosenthal R. (1991): *Meta-analytic Procedures for Social Research*, 2nd ed. Sage Publications, Newbury Park CA. Cyt. za: A. Field (2009): *Discovering Statistics Using SPSS*, 3rd ed. Sage Publications, Newbury Park.

Rosnow R. L., Rosenthal R. (2005): *Beginning Behavioural Research: A Conceptual Primer*, 5th ed. Pearson/Prentice Hall, Englewood Cliffs. Cyt. za: A. Field (2009): *Discovering Statistics Using SPSS*, 3rd ed. Sage Publications, Thousand Oaks.

Royall R. (2000a): *On the Probability of Observing Misleading Statistical Evidence (with Comments).* „Journal of the American Statistical Association", 95, 451, s. 760–780.

Royall R. (2000b): *Statistical Evidence. A Likelihood Paradigm*. Chapman and Hall/CRC, Boca Raton.

Royston P. (1982): *An Extension of Shapiro and Wilks's W Test for Normality to Large Samples.* „Applied Statistics", 31, s. 115–124.

Royston P. (1983): *A Simple Method for Evaluating the Shapiro-Francia W Test of Non-normality*. „Statistician", 32, s. 297–300.

Satten G. A., Kupper L. L. (1990): *Sample Size Requirements for Interval Estimation on the Odds Ratio.* „American Journal of Epidemiology", 131, 1, s. 177–184.

Savage L. J. (1954): *The Foundations of Statistics.* John Wiley and Sons, New York.

Savage L. J. (1958): *Recent Tendencies in the Foundations of Statistics.* Invited address to Section VI of the International Congress of Mathematicians, Edinburgh, 14–21 August 1958.

Schoenfeld D. A. (1983): *Sample-Size Formula for the Proportional-Hazards Regression Model.* „Biometrics", 39, s. 499–503.

Sedlmeier P., Gigerenzer G. (1989): *Do Studies of Statistical Power Have an Effect on the Power of Studies*. „Psychological Bulletin", 105, 2, s. 309–316.

Shapiro S. S., Francia R. S. (1972): *An Approximate Analysis of Variance Test for Normality*. „Journal of the American Statistical Association", 67, s. 215–216.

Shapiro S. S., Wilk M. B. (1965): *An Analysis of Variance Test for Normality (Complete Samples).* „Biometrika", 52, s. 591–611.

Silvey S. D. (1978): *Wnioskowanie statystyczne*. Państwowe Wydawnictwo Naukowe, Warszawa.

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, s. 1–18.

Spiegelhalter D. J., Abrams K. R., Myles J. P. (2004): *Bayesian Approaches to Clinical Trials and Health-Care Evaluation.* John Wiley and Sons, Chichester.

Starek A., Szabla J., Szymczak W., Zapór L. (2006): *Comparison of Acute Haematotoxicity of 2-mthoxyethanol, 2-ethoxyethanol and 2-butoxyethanol in Male Rats*. „Acta Toxicologica", 14, 1–2, s. 63–71.

Sterne J. A. C. (2002): *Teaching Hypothesis Tests – Time for Significant Change?* „Statistics in Medicine", 21, 7, s. 985–994.

Szymczak W. (1999): *Ryzyko choroby nowotworowej a narażenie na substancje chemiczne.* Instytut Medycyny Pracy, Łódź.

Szymczak W. (2018): *Podstawy statystyki dla psychologów*, wydanie trzecie rozszerzone. Wydawnictwo Difin, Warszawa.

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

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 Bill. (2007): *The Nature of Statistical Evidence.* Springer, New York.

Thompson Bruce. (1994): *The Concept of Statistical Significance Testing.* „Practical Assessment, Research and Evaluation", 4, 5, s. 1–3.

Trafimow D. (2003): *Hypothesis Testing and Theory Evaluation at the Boundaries: Surprising Insights from Bayes' Theorem.* „Psychological Review", 110, 3, s. 526–535.

US EPA (1984) (U.S. Environmental Protection Agency): *Health Effects Assessment for Benzo*[*a*]*pyrene.* EPA 540/1-86-022. Environmental Criteria and Assessment Office, Cincinnati.

US EPA (1988) (U.S. Environmental Protection Agency): *Recommendations for and Documentation of Biological Values for Use in Risk Assessment.* EPA 600/6-87-008. Office of Health and Environmental Assessment, Cincinnati.

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

Volker M. A. (2006): *Reporting Effect Size Estimates in School Psychology Research.* „Psychology in the Schools", 43, 6, s. 653–672.

Vorapongsathorn T., Taejaroenkul S., Viwatwongkasem C. (2004): *A Comparison of Type I Error and Power of Bartlett's Test, Levene's Test and Cochran's Test under Violation of Assumptions*. „Songklanakarin Journal of Science and Technology", 26, 4, s. 537–547.

Wasserman L. (2006): *All of Nonparametric Statistics.* Springer, New York.

Welch B. L. (1951): *On the Comparison of Several Mean Values: An Alternative Approach*, „Biometrika", 38, s. 330–336.

Whittemore A. S. (1981): *Sample Size for Logistic Regression with Small Response Probability.* „Journal of the American Statistical Association", 76, 373, s. 27–32.

Wilcox R. R. (2002): *Understanding the Practical Advantages of Modern ANOVA Methods*. „Journal of Clinical Child and Adolescent Psychology", 31, 3, s. 399–412.

Williams R. H., Zimmerman D. W. (1989): *Statistical Power Analysis and Reliability of Measurement*. „Journal of General Psychology", 116, 4, s. 359–369.

Winer (1971): *Statistical Principles in Experimental Design*, 2nd ed. McGraw-Hill, Tokyo.

Winer B. J., Brown D. R., Michels K. M. (1991): *Statistical Principles in Experimental Design*, 3rd ed. McGraw-Hill, Boston.

Woolson R. F. (1987): *Statistical Methods for the Analysis of Biomedical Data.* John Wiley and Sons, New York.

Wuensch K. L. (2015): *Nonparametric Effect Size Estimators*. Materiały dydaktyczne. East Carolina University Department of Psychology, http://core.ecu.edu/psyc/wuenschk/docs30/Nonparametric-EffectSize.pdf (dostęp: 16.06.2015).

Young G. A., Smith R. L. (2005): *Essentials of Statistical Inference.* Cambridge University Press, Cambridge.

Ziliak S. T., McCloskey D. N. (2009): *The Cult of Statistical Significance.* Presented at the Joint Statistical Meetings, August 3rd, Washington.

Zimmerman D. W. (2004): *A Note on Preliminary Tests of Equality of Variances*. „British Journal of Mathematical and Statistical Psychology", 57, s. 173–181.

Zubrzycki S. (1970): *Wykłady z rachunku prawdopodobieństwa i statystyki matematycznej*. Państwowe Wydawnictwo Naukowe, Warszawa.

Załączniki

Załącznik 1

Podstawą wszystkich prezentowanych przykładów będą wyniki rzeczywistych badań naukowych. Wzięto pod uwagę badania Dudka, w których oceniano związek stresu zawodowego ze stanem zdrowia pracowników służb mundurowych (Dudek, 2007), badania Agnieszki Kubot, w których porównywano efektywność trzech terapii w leczeniu „łokcia tenisisty" (Kubot, 2017; Kubot i wsp. 2017) oraz eksperyment toksykologiczny Starka i wsp., w którym oceniano efekty narażenia szczurów na butoksyetanol i oceniano zmienność tych efektów w czasie (Starek i wsp., 2006).

Zmienne występujące w pliku danych badania Dudka

nr_bad: kolejny numer badanego: 1 – 444.
grupa: 1 – straż pożarna, 2 – pracownicy służby więziennej, 3 – policja.
wiek: wiek w latach w chwili badania.
plec: 1 – mężczyźni, 2 – kobiety.
stan_cyw: 1 – kawaler/panna, 2 – żonaty/mężatka, 3 – rozwiedziony(a), 4 – wdowiec/wdowa.
wykszt3: 1 – podstawowe albo zasadnicze, 2 – średnie albo niepełne wyższe, 3 – wyższe.
staz_og: staż pracy ogółem (lata).
subiekt: zmienna zbudowana na podstawie odpowiedzi w kwestionariuszu do subiektywnej oceny pracy; kwestionariusz składa się z 55 pytań i zmienna „subiekt" to suma punktów uzyskana w każdym pytaniu.

Kategorie odpowiedzi:
1 – cecha nie występuje, nie dotyczy mojego stanowiska pracy;
2 – cecha występuje, ale mi nie przeszkadza i nie denerwuje;
3 – czasami mnie to irytuje lub przeszkadza;
4 – dość często mnie to irytuje lub przeszkadza;

5 – irytuje mnie to cały czas w pracy, a nawet denerwuję się z tego powodu w domu.

Zakres zmienności: 55–275. Im niższa wartość zmiennej „subiekt", tym mniejsze odczucie stresogenności pracy.

SOC (poczucie koherencji, ***Sens of Coherence***): zmienna zbudowana na podstawie odpowiedzi na 29 pytań w kwestionariuszu orientacji życiowej (SOC-29). Przy każdym pytaniu podano 7 możliwych odpowiedzi; cyfra 1 w odpowiedzi oznacza najbardziej negatywną reakcję (odczucie), zaś 7 – najbardziej pozytywną reakcję (odczucie) w przypadku badanego problemu. Zakres zmienności: 0–203. Im wyższa wartość zmiennej, tym pozytywniejsza orientacja życiowa.

Siedem kolejnych zmiennych to tzw. profil nastrojów. Składa się z 65 słów opisujących różne odczucia i nastroje, w jakich może znajdować się człowiek. Respondent zaznacza, czy w związku z pracą odczuwał ostatnio stan opisany danym słowem i z jakim natężeniem:
0 – zdecydowanie nie;
1 – raczej nie;
2 – umiarkowanie;
3 – raczej tak;
4 – zdecydowanie tak.

Z tych 65 stanów tworzonych jest 7 następujących profili nastrojów:
wrogosc (wrogość, gniew): zakres 0–48;
zaklopot (zakłopotanie): zakres (–4)–24;
przygneb (przygnębienie): zakres 0–60;
znuzenie (znużenie): zakres 0–28;
zyczliwo (życzliwość): zakres 0–28;
napiecie (napięcie, lęk): zakres (–4)–24;
wigor (wigor, aktywność): zakres 0–32.

Kolejne sześć zmiennych to rezultat kwestionariusza ogólnego stanu zdrowia GHQ 28 Davida Goldberga. Każda pozycja kwestionariusza jest pytaniem o to, czy respondent doświadczył ostatnio jakiegoś określonego objawu lub czy zachowywał się we wskazany w pytaniu sposób. Respondent swoją odpowiedź zaznaczał na skali, do której zastosowano punktację Likerta.

Przykładowo: czy ostatnio czułeś się smutny i ponury?
- mniej niż zwykle – 0;
- nie mniej niż zwykle (tak samo jak zwykle) – 1;
- raczej bardziej niż zwykle – 2;
- znacznie bardziej niż zwykle – 3.

Czy ostatnio udawało Ci się radzić ze wszystkimi swoimi zajęciami?
- lepiej niż zwykle – 0;
- tak samo jak zwykle – 1;
- raczej gorzej niż zwykle – 2;
- znacznie gorzej niż zwykle – 3.

skala GHQ_A (symptomy somatyczne): zakres 0–21.
skala GHQ_B (niepokój, bezsenność): zakres 0–21.
skala GHQ_C (zaburzenia funkcjonowania): zakres 0–21.
skala GHQ_D (symptomy depresji): zakres 0–21.

Im większa wartość zmiennej, tym silniejsze obciążenie respondenta danymi symptomami.
GHQ_suma: suma GHQ_A, GHQ_B, GHQ_C, GHQ_D.

raz_god: liczba godzin pracy w zasadniczym miejscu pracy w ciągu tygodnia.
dod_prac: dodatkowa praca w innym miejscu: 1 – pracuje dodatkowo, 2 – nie pracuje dodatkowo.
wysi_fiz: czy dodatkowa praca związana jest z wysiłkiem fizycznym?

1 – nie;
2 – tak, z małym wysiłkiem;
3 – tak, ze średnim wysiłkiem;
4 – tak, z dużym wysiłkiem;
5 – tak, z bardzo dużym wysiłkiem.

wysifiz3: trzystanowa zmienna charakteryzująca wysiłek fizyczny w dodatkowym miejscu pracy: 1 – brak wysiłku fizycznego; 2 – mały albo średni wysiłek fizyczny; 3 – duży albo bardzo duży wysiłek fizyczny.
czyn_wyp: liczba godzin poświęcanych tygodniowo na czynny wypoczynek wymagający bardzo dużego (regularny trening, udział w zawodach) lub dużego (bieganie, siłownia, pływanie, gra w piłkę, biegi narciarskie, taniec dyskotekowy) wysiłku fizycznego.
wypoczy1: liczba godzin poświęcanych tygodniowo na czynny wypoczynek o małej intensywności (np. spacery).
bier_wyp: liczba godzin poświęcanych tygodniowo na bierny wypoczynek.
hobby: czy w ciągu ostatniego tygodnia znalazł(a) Pan(i) czas na to, by robić to, co Pan(i) chciał(a) i lubił(a)?

1 – tak;
2 – nie.

sen: ile godzin powinien Pan(i) spać, aby czuć się wyspanym?
brak_snu: ile razy w ciągu tygodnia zdarza się Pan(i) spać mniej niż Pan(i) powinien (powinna)?
palenie3:

0 – nie pali i nie palił;
1 – pali obecnie;
2 – były palacz.

alko_rok: oszacowana, na podstawie częstości picia piwa, wina i wódki oraz ilości tych napojów wypijanych jednorazowo, ilość czystego spirytusu wypijana w ciągu roku.
nadc_fa: stwierdzona choroba nadciśnieniowa.
ukl_kraz: stwierdzona niedokrwienna choroba serca lub inna choroba serca.

ukl_odde: stwierdzona choroba układu oddechowego (np. przewlekły nieswoisty nieżyt oskrzeli).
ukl_nerw: stwierdzona choroba układu nerwowego (np. choroby obwodowego układu nerwowego).
ukl_poka: stwierdzona choroba układu pokarmowego (nieżyt żołądka, choroba wrzodowa żołądka lub dwunastnicy, choroby wątroby, choroby trzustki).
ukl_ruch: stwierdzona choroba układu ruchu (dolegliwości ze strony kręgosłupa w odcinku szyjnym, piersiowym, lędźwiowo-krzyżowym, kończyn górnych, kończyn dolnych).
ukl_dokr: stwierdzona choroba układu dokrewnego albo choroba przemiany materii (np. cukrzyca).
alergie: stwierdzona choroba o podłożu alergicznym (np. astma oskrzelowa, pyłkowica, zmiany skórne).

Zmienne od „nadc_fa" do „alergie" to zmienne dwustanowe, tj. mogące przyjmować tylko dwie wartości. Nazywane są także zmiennymi zero-jedynkowymi (zero oznacza niewystępowanie badanego zjawiska czy stanu, a jedynka oznacza jego występowanie) albo zmiennymi dychotomicznymi.

stazdrfa: zmienna dychotomiczna charakteryzująca stan zdrowia: 0 – nie stwierdzono chorób przewlekłych, 1 – stwierdzono u pacjenta przynajmniej jedną chorobę przewlekłą.
cholest: cholesterol całkowity w surowicy (mg/dl).
HDL: cholesterol HDL w surowicy (mg/dl).
LDL: cholesterol LDL w surowicy (mg/dl).
TG: trójglicerydy w surowicy (mg/dl).
cukier: poziom cukru w surowicy na czczo (ml/dl).
ciezar: masa ciała (kg).
wzrost: wzrost (cm).
BMI: indeks masy ciała (Body Mass Index) (masa ciała w kg podzielona przez kwadrat wzrostu wyrażonego w metrach).

Wybrane zmienne spośród występujących w pliku danych badania Kubot

Kwestionariusz **DASH** jest 30-punktowym kwestionariuszem samooceny funkcji kończyny górnej. Zmienna ta była oceniana w trzech punktach czasowych:
DASH_0 – ocena przed terapią;
DASH_3 – ocena bezpośrednio po terapii;
DASH_8 – ocena po ośmiu tygodniach od zakończenia terapii.

Skala VAS – Wizualna Analogowa Skala Bólu (*Visual Analogue Scale*) – pozwala badanemu „opisać" natężenie odczuwanych dolegliwości. Chory, posługując się jedenastostopniową miarą, określa natężenie bólu, korzystając z wartości numerycznych (gdzie „0" oznacza całkowity brak bólu, natomiast „10" – najsilniejszy

ból możliwy do wyobrażenia). Zmienna VAS też była oceniana w trzech punktach czasowych:
VAS_0 – ocena przed terapią;
VAS_3 – ocena bezpośrednio po terapii;
VAS_8 – ocena po ośmiu tygodniach od zakończenia terapii.

Rodzaj terapii „łokcia tenisisty":
1 – terapia klasyczna, grupę tę potraktowano jako grupę kontrolną;
2 – terapia falą uderzeniową;
3 – terapia ultradźwiękami.

Wybrane zmienne spośród występujących w pliku danych badania Starka i wsp.

Dane z tego eksperymentu były wykorzystane jedynie w przykładzie 4.10. Cztery grupy samców szczurów, po pięć zwierząt w każdej, narażone zostały na 2-butoksyetanol w dawkach: 0; 0,25; 0,50 i 0,75 mmol/kg masy ciała. Oceniano parametry krwi przed narażeniem oraz 4, 11, 18 i 28 dób po narażeniu. W przykładzie wykorzystano parametr RBC (czerwone krwinki).

Załącznik 2

Wyniki jednoczynnikowej analizy wariancji w trzech programach statystycznych: SPSS 24, STATA 13 i SYSTAT 13.

Program SPSS 24

```
ONEWAY DASH_8 BY kod_grupy3
 /STATISTICS DESCRIPTIVES HOMOGENEITY BROWNFORSYTHE WELCH
 /MISSING ANALYSIS
 /POSTHOC=BONFERRONI T3 ALPHA(0.05).
```

Jednoczynnikowa analiza wariancji (ONEWAY)

Uwagi		
Raport sporządzono		18-JUN-2017 17:09:20
Komentarze		
Dane wejściowe	Plik danych	D:\dysk_E\dydaktyka\zajecia16_17\wyklad_metody_statystyczne\egzamin\Agnieszka_dane_rob.sav
	Roboczy plik danych	ZbiórDanych1
	Filtr	<brak>
	Waga	<brak>
	Podział na podzbiory	<brak>
	Liczba obserwacji w roboczym pliku danych	120
Traktowanie braków danych	Definicja braków danych	Wartości zdefiniowane przez użytkownika jako braki danych są traktowane jako braki danych
	Użycie obserwacji	Statystyki obliczane są na podstawie obserwacji, które nie mają braków danych w żadnej ze zmiennych użytych w danej analizie
Komenda		ONEWAY DASH_8 BY kod_grupy3 /STATISTICS DESCRIPTIVES HOMOGENEITY BROWNFORSYTHE WELCH /MISSING ANALYSIS /POSTHOC=BONFERRONI T3 ALPHA(0.05).
Zasoby	Czas procesora	00:00:00,00
	Czas wykonania	00:00:00,02

Statystyki opisowe								
DASH_8								
	N	Średnia	Odchylenie standardowe	Błąd standardowy	95% przedział ufności dla średniej		Minimum	Maksimum
					dolna granica	górna granica		
,00	60	42,1370	25,37181	3,27549	35,5827	48,6912	,00	90,15
1,00	30	14,4819	12,83932	2,34413	9,6876	19,2762	,00	43,97
2,00	30	32,4094	19,47637	3,55588	25,1368	39,6820	,00	69,82
Ogółem	120	32,7913	24,09070	2,19917	28,4367	37,1459	,00	90,15

Test jednorodności wariancji			
DASH_8			
Test Levene'a	*df*1	*df*2	Istotność
10,737	2	117	,000

Jednoczynnikowa ANOVA					
DASH_8					
	Suma kwadratów	*df*	Średni kwadrat	*F*	Istotność
Między grupami	15 301,907	2	7650,954	16,651	,000
Wewnątrz grup	53 761,147	117	459,497		
Ogółem	69 063,055	119			

Mocne testy równości średnich				
DASH_8				
	Statystyka[a]	*df*1	*df*2	Istotność
Welch	25,578	2	69,872	,000
Brown-Forsythe	20,962	2	105,026	,000

[a] Rozkład *F* asymptotyczny.

Testy *post hoc*

Porównania wielokrotne							
Zmienna zależna: DASH_8							
Testy	(I) kod_ grupy3	(J) kod_ grupy3	Różnica średnich (I–J)	Błąd standardowy	Istotność	95% przedział ufności	
						dolna granica	górna granica
Test Bonferroniego	,00	1,00	27,65508*	4,79321	,000	16,0129	39,2973
		2,00	9,72760	4,79321	,134	–1,9146	21,3698
	1,00	,00	–27,65508*	4,79321	,000	–39,2973	–16,0129
		2,00	–17,92749*	5,53472	,005	–31,3707	–4,4843
	2,00	,00	–9,72760	4,79321	,134	–21,3698	1,9146
		1,00	17,92749*	5,53472	,005	4,4843	31,3707
Test Dunnetta T3	,00	1,00	27,65508*	4,02787	,000	17,8579	37,4523
		2,00	9,72760	4,83458	,136	–2,0761	21,5313
	1,00	,00	–27,65508*	4,02787	,000	–37,4523	–17,8579
		2,00	–17,92749*	4,25902	,000	–28,4336	–7,4214
	2,00	,00	–9,72760	4,83458	,136	–21,5313	2,0761
		1,00	17,92749*	4,25902	,000	7,4214	28,4336

* Różnica średnich jest istotna na poziomie 0,05.

```
EXAMINE VARIABLES=DASH_8 BY kod_grupy3
 /PLOT NPPLOT
 /STATISTICS DESCRIPTIVES
 /CINTERVAL 95
 /MISSING PAIRWISE
 /NOTOTAL.
```

Eksploracja (EXAMINE)

Uwagi		
Raport sporządzono		19-JUN-2017 18:39:03
Komentarze		
Dane wejściowe	Plik danych	D:\dysk_E\dydaktyka\wnioskowanie_statystyczne\czerwiec_2017\porownania_progamow\A_rob.sav
	Roboczy plik danych	ZbiórDanych1
	Filtr	<brak>
	Waga	<brak>
	Podział na podzbiory	<brak>
	Liczba obserwacji w roboczym pliku danych	120
Traktowanie braków danych	Definicja braków danych	Wartości zdefiniowane przez użytkownika jako braki danych dla zmiennych zależnych są traktowane jako braki danych.
	Użycie obserwacji	Statystyki obliczane są na podstawie obserwacji, które nie mają braków danych w analizowanych zmiennych zależnych lub czynnikach.
Komenda		EXAMINE VARIABLES=DASH_8 BY kod_grupy3 /PLOT NPPLOT /STATISTICS DESCRIPTIVES /CINTERVAL 95 /MISSING PAIRWISE /NOTOTAL.
Zasoby	Czas procesora	00:00:01,99
	Czas wykonania	00:00:01,61

kod_grupy3

Informacja o analizowanych danych							
	kod_ grupy3	Obserwacje					
		uwzględnione		wykluczone		ogółem	
		N	procent	N	procent	N	procent
DASH_8	0	60	100,0%	0	0,0%	60	100,0%
	1	30	100,0%	0	0,0%	30	100,0%
	2	30	100,0%	0	0,0%	30	100,0%

Statystyki opisowe (DESCRIPTIVES)					
	kod_grupy3			Statystyka	Błąd standardowy
DASH_8	0	Średnia		42,136957	3,2754871
		95% przedział ufności dla średniej	dolna granica	35,582722	
			górna granica	48,691191	
		5% średnia obcięta		41,813285	
		Mediana		44,820000	
		Wariancja		643,729	
		Odchylenie standardowe		25,3718140	
		Minimum		,0000	
		Maksimum		90,1500	
		Rozstęp		90,1500	
		Rozstęp ćwiartkowy		41,1633	
		Skośność		,026	,309
		Kurtoza		–1,109	,608
	1	Średnia		14,481873	2,3441277
		95% przedział ufności dla średniej	dolna granica	9,687594	
			górna granica	19,276153	
		5% średnia obcięta		13,712322	
		Mediana		10,775000	
		Wariancja		164,848	
		Odchylenie standardowe		12,8393162	
		Minimum		,0000	
		Maksimum		43,9650	
		Rozstęp		43,9650	
		Rozstęp ćwiartkowy		19,3960	
		Skośność		,733	,427
		Kurtoza		–,465	,833

DASH_8	2	Średnia		32,409360	3,5558833
		95% przedział ufności dla średniej	dolna granica	25,136762	
			górna granica	39,681958	
		5% średnia obcięta		32,179270	
		Mediana		34,913000	
		Wariancja		379,329	
		Odchylenie standardowe		19,4763750	
		Minimum		,0000	
		Maksimum		69,8200	
		Rozstęp		69,8200	
		Rozstęp ćwiartkowy		33,1915	
		Skośność		–,035	,427
		Kurtoza		–,935	,833

Testy normalności rozkładu							
	kod_grupy3	Kołmogorow-Smirnow[a]			Shapiro-Wilk		
		Statystyka	*df*	Istotność	Statystyka	*df*	Istotność
DASH_8	0	,108	60	,081	,958	60	,039
	1	,156	30	,060	,908	30	,013
	2	,105	30	,200*	,962	30	,339

* Dolna granica rzeczywistej istotności.

[a] Z poprawką istotności Lillieforsa.

Program STATA 13

```
------------------------------------------------------------------------------
. anova DASH_8 kod_grupy3

      Number of obs =   120   R-squared   = 0.2216
      Root MSE   = 21.4359   Adj R-squared = 0.2083

  Source | Partial SS  df    MS      F   Prob > F
-----------+----------------------------------------------------
   Model | 15301.9074   2  7650.9537   16.65   0.0000
     |
kod_grupy3 | 15301.9074   2  7650.9537   16.65   0.0000
```

```
           |
  Residual | 53761.1471  117 459.496984
-----------+----------------------------------------------------
     Total | 69063.0545  119 580.361803
```

. oneway DASH_8 kod_grupy3, bonferroni scheffe sidak tabulate

```
    |    Summary of DASH_8
kod_grupy3 |    Mean  Std. Dev.    Freq.
------------+------------------------------------
    0 | 42.136957  25.371814     60
    1 | 14.481873  12.839316     30
    2 | 32.40936  19.476375     30
------------+------------------------------------
  Total | 32.791287   24.0907     120

          Analysis of Variance
 Source      SS     df   MS      F   Prob > F
------------------------------------------------------------------------
Between groups   15301.9074   2  7650.9537   16.65   0.0000
 Within groups   53761.1471  117  459.496984
------------------------------------------------------------------------
 Total     69063.0545  119  580.361803
```

Bartlett's test for equal variances: chi2(2) = 15.2007 Prob>chi2 = 0.001

```
         Comparison of DASH_8 by kod_grupy3
              (Bonferroni)
Row Mean-|
Col Mean |     0     1
---------+----------------------
   1 |  -27.6551
     |   0.000
     |
   2 |  -9.7276  17.9275
     |   0.134   0.005

         Comparison of DASH_8 by kod_grupy3
              (Scheffe)
Row Mean-|
Col Mean |     0     1
---------+----------------------
   1 |  -27.6551
     |   0.000
     |
   2 |  -9.7276  17.9275
     |   0.132   0.007
```

```
        Comparison of DASH_8 by kod_grupy3
                (Sidak)
Row Mean-|
Col Mean |     0     1
---------+----------------------
   1 |  -27.6551
     |   0.000
     |
   2 |  -9.7276  17.9275
     |   0.128   0.005
```

```
. estat summarize

 Estimation sample anova       Number of obs =  120

 ------------------------------------------------------------
  Variable |    Mean   Std. Dev.    Min    Max
 -------------+----------------------------------------------
   DASH_8 | 32.79129   24.0907     0   90.15
      |
 kod_grupy3 |
     1 |     .25   .4348283     0     1
     2 |     .25   .4348283     0     1
 ------------------------------------------------------------
```

```
. robvar DASH_8, by(kod_grupy3)

     |    Summary of DASH_8
 kod_grupy3 |    Mean  Std. Dev.    Freq.
------------+------------------------------------
    0 |  42.136957  25.371814     60
    1 |  14.481873  12.839316     30
    2 |  32.40936  19.476375     30
------------+------------------------------------
  Total |  32.791287   24.0907     120
W0 = 10.736956  df(2, 117)   Pr > F = 0.00005234
W50 = 10.048848  df(2, 117)   Pr > F = 0.00009389
W10 = 10.923764  df(2, 117)   Pr > F = 0.00004471
```

```
. median DASH_8, by(kod_grupy3) exact medianties(below)

Median test
Enumerating sample-space combinations:
stage 3: enumerations = 1
stage 2: enumerations = 12
stage 1: enumerations = 0

  Greater |
 than the |      kod_grupy3
```

```
 median |    0     1     2 |  Total
-----------+---------------------------------+----------
    no |   22    26    12 |   60
   yes |   38     4    18 |   60
-----------+---------------------------------+----------
 Total |   60    30    30 |  120

    Pearson chi2(2) = 21.6000  Pr = 0.000
     Fisher's exact =         0.000

. by kod_grupy3, sort : swilk DASH_8


-> kod_grupy3 = 0

        Shapiro-Wilk W test for normal data

 Variable | Obs    W      V     z    Prob>z
-------------+--------------------------------------------------
  DASH_8 |  60  0.95874   2.243   1.741  0.04086

-> kod_grupy3 = 1

        Shapiro-Wilk W test for normal data

 Variable | Obs    W      V     z    Prob>z
-------------+--------------------------------------------------
  DASH_8 |  30  0.90497   3.020   2.286  0.01114
-------------------------------------------------------------------------------------------
-> kod_grupy3 = 2

        Shapiro-Wilk W test for normal data

 Variable | Obs    W      V     z    Prob>z
-------------+--------------------------------------------------
  DASH_8 |  30  0.96225   1.200   0.377  0.35321
```

. by kod_grupy3, sort : sfrancia DASH_8, boxcox

```
-------------------------------------------------------------------------------------------
-> kod_grupy3 = 0

       Shapiro-Francia W' test for normal data

 Variable | Obs    W'     V'    z    Prob>z
-------------+--------------------------------------------------
  DASH_8 |  60  0.96837   1.894   1.236  0.10815
-------------------------------------------------------------------------------------------
-> kod_grupy3 = 1
```

```
      Shapiro-Francia W' test for normal data

 Variable | Obs    W'     V'    z    Prob>z
------------+-------------------------------------------------
  DASH_8 |  30  0.94177   2.044   1.311  0.09488
-------------------------------------------------------------------------------
-> kod_grupy3 = 2

      Shapiro-Francia W' test for normal data

 Variable | Obs    W'     V'    z    Prob>z
------------+-------------------------------------------------
  DASH_8 |  30  0.97273   0.957  -0.081  0.53242
```

. by kod_grupy3, sort : sfrancia DASH_8

```
-------------------------------------------------------------------------------
-> kod_grupy3 = 0

      Shapiro-Francia W' test for normal data
 Variable | Obs    W'     V'    z    Prob>z
------------+-------------------------------------------------
  DASH_8 |  60  0.96837   1.903   1.231  0.10923
-------------------------------------------------------------------------------
-> kod_grupy3 = 1

      Shapiro-Francia W' test for normal data

 Variable | Obs    W'     V'    z    Prob>z
------------+-------------------------------------------------
  DASH_8 |  30  0.94177   2.054   1.318  0.09372
-------------------------------------------------------------------------------
-> kod_grupy3 = 2

      Shapiro-Francia W' test for normal data

 Variable | Obs    W'     V'    z    Prob>z
------------+-------------------------------------------------
  DASH_8 |  30  0.97273   0.962  -0.072  0.52854
-------------------------------------------------------------------------------
```

Program SYSTAT 13

Number of Variables	:	67
Number of Cases	:	120

SYSTAT Rectangular file D:\dysk_E\dydaktyka\wnioskowanie_statystyczne\czerwiec_2017\porownania_progamow\A_data.syz,
Created data file Sun Jun 18 16:05:08 2017 containing variables:

GRUPA$	KOD_GRUPY	WIEK	PLEC	MASA	WZROST
BMI	ZAWOD	STAZ	DASH_0	DASH_3	DASH_8
DASHPRA-CA_0	DASHPRA-CA_3	DASHPRA-CA_8	WSK_ DASH_0	WSK_ DASH_3	WSK_ DASH_8
WSK_ DASH_0_3	WSK_ DASH_3_8	WSK_DASH-PRACA_0	WSK_DASH-PRACA_3	WSK_DASH-PRACA_8	WSK_DASH_ PRACA_-0_3
WSK_DASH_ PRACA_-3_8	VAS_0	VAS_3	VAS_8	VAS_U_0	VAS_U_3
VAS_U_8	MILL_0	MILL_3	MILL_8	MILL_0_3	MILL_3_8
PROBA_K_0	PROBA_K_3	PROBA_K_8	PROBA-_K_0_3	PROBA-_K_3_8	THOM-SON_0
TOMSON_3	THOM-SON_8	THOM-SON_0_3	THOM-SON_3_8	NASILENIE_ BOLU_-0	CZESTOTLI-WOSC_B--OLU_0
UZYWANIE_ LEKOW_-0	OGR_ SPRAWN_0	NASILENIE_ BOLU_-3	CZESTOTLI-WOSC_B--OLU_3	UZYWANIE_ LEKOW_-3	OGR_ SPRAWN_3
NASILENIE_ BOLU_-8	CZESTOTLI-WOSC_B--OLU_8	UZYWANIE_ LEKOW_-8	OGR_ SPRAWN_8	NASILENIE_ BOLU_-0_3	NASILENIE_ BOLU_-3_8
CZESTOTLI-WOSC_B--OLU_0_3	CZESTOTLI-WOSC_B--OLU_3_8	UZYWANIE_ LEKOW_-0_3	UZYWANIE_ LEKOW_-3_8	OGR_ SPRAWNO-SCI_-0_3	OGR_ SPRAWNO-SCI_-3_8
KOD_GRU-PY3					

> REM -- Following commands were produced by the ANOVA dialog:
> ANOVA
> DEPEND DASH_8
> SUBCAT KOD_GRUPY3 / EFFECT
> COVAR
> ESTIMATE / NTEST = {KS, AD, SW} HTEST = LEVENE SS = TYPE3

▼Analysis of Variance

Effects coding used for categorical variables in model.
The categorical values encountered during processing are

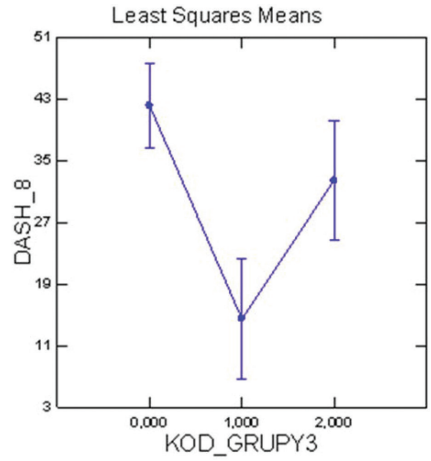
Variables	Levels		
KOD_GRUPY3 (3 levels)	0,000	1,000	2,000

Dependent Variable	DASH_8
N	120
Multiple R	0,471
Squared Multiple R	0,222

Estimates of Effects B = $(X'X)^{-1}X'Y$		
Factor	**Level**	**DASH_8**
CONSTANT		29,676
KOD_GRUPY3	0,000	12,461
KOD_GRUPY3	1,000	–15,194

Analysis of Variance					
Source	**Type III SS**	**df**	**Mean Squares**	**F-Ratio**	**p-Value**
KOD_GRUPY3	15 301,907	2	7 650,954	16,651	0,000
Error	53 761,147	117	459,497		

Least Squares Means				
Factor	**Level**	**LS Mean**	**Standard Error**	**N**
KOD_GRUPY3	0,000	42,137	2,767	60,000
KOD_GRUPY3	1,000	14,482	3,914	30,000
KOD_GRUPY3	2,000	32,409	3,914	30,000



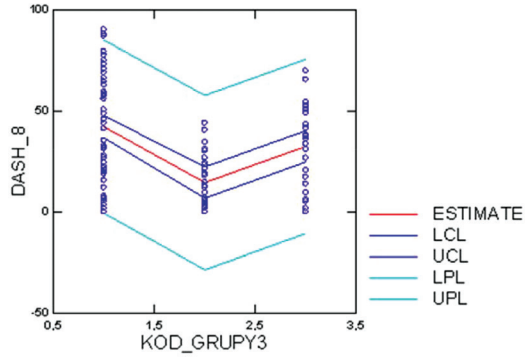
Levene's Test for Homogeneity of Variances		
	Test Statistic	**p-Value**
Based on Mean	10,737	0,000
Based on Median	10,049	0,000

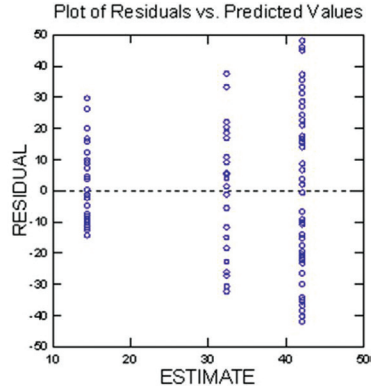
Test for Normality		
	Test Statistic	**p-Value**
K-S Test (Lilliefors)	0,063	0,267
Shapiro-Wilk Test	0,985	0,205
Anderson-Darling Test	0,402	> 0,15*

* The p-value cannot be precisely computed.

Durbin-Watson D-Statistic	1,767
First Order Autocorrelation	0,110

Information Criteria	
AIC	1 081,123
AIC (Corrected)	1 081,471
Schwarz's BIC	1 092,273





> REM -- End of commands from the ANOVA dialog
> REM -- Following commands were produced by the ANOVAHYPO dialog:
> HYPOTHESIS

▼Hypothesis Tests

Test for effect called: KOD_GRUPY3

Contrast Estimate				
Hypothesis	**Estimate(AB)**	**Standard Error**	**95% Confidence Interval**	
			lower	**upper**
1	12,461	2,609	7,294	17,628
2	–15,194	3,059	–21,253	–9,135

Inverse Contrast $A(X'X)^{-1}A'$		
	1	**2**
1	0,015	
2	–0,007	0,020

Test for Contrast					
Source	**SS**	**df**	**Mean Squares**	**F-Ratio**	**p-Value**
1	10 480,986	1	10 480,986	22,810	0,000
2	11 333,295	1	11 333,295	24,665	0,000

Test of Hypothesis					
Source	**SS**	**df**	**Mean Squares**	**F-Ratio**	**p-Value**
KOD_GRUPY3	15 301,907	2	7 650,954	16,651	0,000
Error	53 761,147	117	459,497		

> EFFECT KOD_GRUPY3
> TEST / CONFI = 0.95
> REM -- End of commands from the ANOVAHYPO dialog
> REM -- Following commands were produced by the ANOVAPOST dialog:
> HYPOTHESIS

▼Hypothesis Tests

Post Hoc Test of DASH_8
Using least squares means.
Using separate variances error terms.

Dunnett's T3 Test					
KOD_GRUPY3(i)	**KOD_GRUPY3(j)**	**Difference**	**p-Value**	**95% Confidence Interval**	
				Lower	**Upper**
0,000	1,000	27,655	0,000	17,858	37,452
0,000	2,000	9,728	0,136	–2,076	21,531
1,000	2,000	–17,927	0,000	–28,434	–7,421

> POST KOD_GRUPY3 / T3
> TEST / CONFI = 0.95
> REM -- End of commands from the ANOVAPOST dialog

Słowniczek

Błąd drugiego rodzaju – to błąd powstający podczas testowania hipotez statystycznych, polegający na przyjęciu fałszywej hipotezy zerowej. Oczywiście nie wiemy, czy hipoteza zerowa jest obiektywnie fałszywa.

Błąd pierwszego rodzaju – to błąd popełniany podczas testowania hipotez statystycznych, polegający na odrzuceniu prawdziwej hipotezy zerowej. Oczywiście nie wiemy, czy hipoteza zerowa jest obiektywnie prawdziwa.

Funkcja mocy testu – jest to funkcja parametru θ (parametr charakteryzujący rozkład prawdopodobieństwa) określająca prawdopodobieństwo odrzucenia hipotezy zerowej. Prawdopodobieństwo to zależy więc od wartości parametru θ.

Hipoteza alternatywna – „druga" z hipotez w zagadnieniu testowania w teorii Neymana-Pearsona. W praktyce hipoteza alternatywna jest prawie zawsze hipotezą złożoną, w przeciwieństwie do hipotezy zerowej, która praktycznie zawsze jest hipotezą prostą.

Hipoteza prosta – hipotezę statystyczną precyzującą wartości wszystkich analizowanych parametrów nazywamy hipotezą prostą. Sformułowanie to jest dość nieprecyzyjne. Czy, przykładowo, hipoteza H_o: $\mu_1 = \mu_2$ precyzuje wartości oczekiwane μ_1 i μ_2? Określa relację między nimi, ale traktuje się ją jako hipotezę prostą. Analogiczna sytuacja pojawia się w przypadku hipotezy nieparametrycznej H_o: $F_1 = F_2$, gdzie F_1 i F_2 są rozkładami prawdopodobieństwa (dystrybuantami) pewnej badanej cechy. Ta hipoteza też jest uważana za prostą.

Hipoteza statystyczna – to każde przypuszczenie dotyczące rozkładu prawdopodobieństwa badanej zmiennej losowej.

Hipoteza zerowa – „pierwsza" z hipotez w zagadnieniu testowania w teorii Neymana-Pearsona. Praktycznie zawsze jest to hipoteza prosta. Badaczowi najczęściej zależy na odrzuceniu hipotezy zerowej, tj. uznaniu jej za fałszywą. Takie traktowanie hipotezy zerowej wprowadza mnóstwo zakłóceń do teorii Neymana-Pearsona.

Hipoteza złożona – hipotezą złożoną możemy nazwać każdą hipotezę statystyczną, która nie będzie hipotezą prostą (co prawda jest to pewien wybieg).

Moc testu statystycznego – w pewnym przybliżeniu można ją traktować jako prawdopodobieństwo odrzucenia fałszywej hipotezy zerowej, czyli moc testu = 1 – β, gdzie β jest prawdopodobieństwem popełnienia błędu drugiego rodzaju. Jednak problem polega na tym, że wartość β jest zależna od konkretnej hipotezy zerowej. Zatem moc testu jest pewną funkcją, a nie liczbą. Niekiedy ta moc testu nazywana bywa mocą *a priori*.

Obserwowana moc testu – nie spotkałem definicji obserwowanej mocy testu. W podręcznikach pisanych przez badaczy w naukach społecznych nazywa się to niekiedy mocą testu *a posteriori*, czyli po zaobserwowaniu próby. Skoro moc *a priori* testu zależy od konkretnej hipotezy alternatywnej (jednego ustalonego punktu w złożonej hipotezie alternatywnej), to nie potrafimy obliczyć mocy *a posteriori*. Niektórzy autorzy twierdzą, że to, co jest obliczane w programach statystycznych jako obserwowana moc testu to pewna uśredniona wartość (czego?).

Paradygmat w statystyce – przez paradygmat w statystyce będziemy rozumieć określoną teorię – czy to testowania hipotez statystycznych, czy estymacji. Możemy spotkać się z teorią częstotliwościową, teorią bayesowską, teorią wiarygodnościową lub kilkoma innymi.

Podejście bayesowskie (paradygmat bayesowski) – z punktu widzenia filozofii statystyki paradygmat bayesowski opiera się na interpretacji prawdopodobieństwa jako racjonalnej warunkowej mierze niepewności, która jest blisko związana z rozumieniem słowa „prawdopodobieństwo" w języku potocznym (Bernardo, 2011). Z praktycznego punktu widzenia w podejściu bayesowskim każdy parametr charakteryzujący rozkład jest traktowany jako zmienna losowa o określonym, znanym rozkładzie prawdopodobieństwa.

Podejście wiarygodnościowe (paradygmat wiarygodnościowy) – w podejściu wiarygodnościowym podstawą wnioskowania jest funkcja wiarygodności, wielowymiarowa funkcja parametrów rozkładu. Wartości funkcji wiarygodności pozwalają porównywać tworzone modele i wybierać najlepszy (według kryterium wiarygodnościowego). Metoda ta jest podstawowa przy szacowaniu współczynników regresji w modelach regresji logistycznej.

Poziom istotności testu statystycznego – przez poziom istotności testu statystycznego będziemy rozumieli górne ograniczenie prawdopodobieństwa błędu pierwszego rodzaju w analizowanym zagadnieniu.

Test hipotezy statystycznej – przez test hipotezy statystycznej będziemy rozumieli pewną procedurę umożliwiającą nam podjęcie decyzji o prawdziwości albo fałszywości odpowiedniej hipotezy statystycznej. Jednak test pozwala jedynie na podjęcie decyzji o prawdziwości albo fałszywości hipotezy, natomiast nie pozwala obiektywnie rozstrzygnąć, czy hipoteza będąca przedmiotem naszego zainteresowania jest prawdziwa albo fałszywa.

Wielkość efektu – grupa różnych mierników występujących w opracowaniach badaczy w naukach społecznych, ekonomicznych, naukach o zarządzaniu, która w pewnym sensie ma doprecyzować wynik testowania hipotez statystycznych. W skrajnych przypadkach przy ich pomocy autorzy próbują zastąpić testowanie. W książkach i artykułach pisanych przez matematyków i statystyków teoretyków pojęcie to nie występuje – jako niemające podstaw teoretycznych.

Indeks stosowanych terminów

Spis tabel i rycin

Tabele

Ryciny