POZIOM ISTOTNOŚCI I GRANICA ROZSĄDKU – PROBLEM PORÓWNAŃ WIELOKROTNYCH W BADANIACH NAUKOWYCH NA PRZYKŁADACH Z ZAKRESU BIOLOGII BEHAWIORALNEJ CZŁOWIEKA

Dariusz Danel, Instytut Immunologii i Terapii Doświadczalnej im. Ludwika Hirszfelda, Polska Akademia Nauk, Wrocław

Wprowadzenie

Jednym z najważniejszych zastosowań statystyki jest uzyskiwane informacji o populacji generalnej na podstawie próby. Cel ten osiągany jest poprzez weryfikację hipotez statystycznych za pomocą odpowiednich schematów postępowań nazywanych testami statystycznymi. Stanisz (2006) definiuje 5 ogólnych etapów w procesie testowania statystycznego:

1. formułowanie hipotezy zerowej (H_0) oraz, odpowiadającej jej, hipotezy alternatywnej (H_1),
2. określenie poziomu istotności statystycznej,
3. wyliczenie wartości odpowiednio dobranego testu statystycznego na podstawie danych z próby,
4. porównanie otrzymanych wartości testu z wartościami krytycznymi ustalonymi dla danego poziomu istotności,
5. podjęcie decyzji o nieodrzuceniu H_0 lub jej odrzuceniu i przyjęciu H_1 na określonym poziomie istotności.

Nawet pobieżny przegląd powyższych punktów zwraca uwagę na jeden powtarzający się termin o fundamentalnym znaczeniu dla procesu wnioskowania statystycznego. Mowa o poziomie istotności statystycznej oznaczanym jako α. Termin ten definiowany jest jako prawdopodobieństwo uzyskania takiego wyniku testu statystycznego, który uprawnia do odrzucenia H_0 i przyjęcia H_1 w sytuacji, gdy H_0 w rzeczywistości jest prawdziwa. Innymi słowy, α oznacza maksymalne, możliwe do zaakceptowania przez badacza, ryzyko pomyłki i odrzucenia poprawnej H_0. Taki błąd w języku statystyki jest definiowany jako błąd I rodzaju i zwykle określony na poziomie $\alpha = 0,05$ lub niższym, np. $\alpha = 0,01$ czy $\alpha = 0,001$. Błędu I rodzaju nie da się uniknąć, bowiem zawsze, choćby z minimalnym prawdopodobieństwem, możliwe jest przypadkowe otrzymanie takiego wyniku, który zasugeruje badaczowi odrzucenie poprawnej hipotezy zerowej. Przykładem może być eksperyment polegający na rzucie monetą, w którym H_0 zakłada, że prawdopodobieństwo otrzymania reszki wynosi $p = 0,5$. Nawet gdy w 10 rzutach nie wyrzucimy ani jednej reszki, to ciągle nie możemy mieć 100% pewności, że uzyskany przez nas wynik nie jest dziełem przypadku. Zgodnie z teorią rozkładu dwumianowego prawdopodobieństwo uzyskania 10 orłów na 10 rzutów rzetelną monetą[13] wynosi dokładnie $p = 0,001$. Wartość ta jest jednocześnie ryzykiem popełnienia błędu I rodzaju, na który eksperymentator musi się zgodzić, odrzucając H_0. Tak więc przy prawdziwej H_0 poziom α można zdefiniować jako prawdopodobieństwo pojawiania się określonego wyniku przez zwyczajny przypadek. Jednocześnie interpretacja taka oznacza, że przy prawdziwej H_0 i ustalonym poziomie istotności $\alpha = 0,05$ oraz przy przeprowadzaniu 100 razy testów statystycznych można spodziewać się pojawienia 5 przypadków, w których testy będą fałszywie istotne statystycznie (tzw. fałszywie pozytywne).

Nieuzasadnione odrzucenie poprawnej hipotezy zerowej nie jest jedynym błędem, na jaki narażony jest badacz w procesie testowania statystycznego. Niejako przeciwnym do błędu I rodzaju jest tzw. błąd II rodzaju. Błąd ten jest popełniany w sytuacji, gdy nieodrzucona jest H_0, a w rzeczywistości prawdziwą jest H_1. Prawdopodobieństwo popełnienia takiego

[13] „Rzetelna moneta" oznacza monetę, w której szansa wyrzucenia reszki lub orła jest taka sama i faktycznie wynosi p=q=0,5.

błędu oznaczane jest jako β. Zwyczajowo przyjmuje się, że akceptowalne prawdopodobieństwo pojawienia się tego błędu wynosi $\beta = 0{,}20$. Wartość taka oznacza, że w 20 przypadkach na 100 nie da się wykryć prawdziwej H_1 i prawdziwy efekt eksperymentalny zostanie przeoczony. Choć w warunkach eksperymentalnych błąd ten można minimalizować poprzez zwiększanie liczebności próby (a przez to dokładności eksperymentu), to jednak sprowadzenie go do zera jest praktycznie niemożliwe, bowiem wymagałoby przebadania całej populacji generalnej.

Oczywistym jest, że oba rodzaje błędów są ze sobą logicznie powiązane. Zmniejszenie poziomu α (ryzyka błędu I rodzaju), czyli zaostrzenie kryteriów zaakceptowania otrzymanego wyniku jako istotnego, spowoduje jednoczesne (choć nie prostoliniowo) zwiększenie ryzyka błędu II rodzaju (tj. poziomu β) i przeoczenia rzeczywiście istniejących różnic czy zależności. Charakterystyczna jest również znacząca dysproporcja obu błędów, na jakie zwyczajowo godzi się badacz. Wynika ona ze sposobu formułowania hipotez. Podczas gdy H_0 jest hipotezą „bezpieczną", której nieodrzucenie nie podważa ogólnie założonych praw czy sądów, to H_1 jest hipotezą konkurencyjną, „odważną", zakładającą istnienie nieznanego do tej pory efektu czy zależności. Tradycyjnie uważa się, że nieodrzucanie nieprawdziwej H_0 jest mniej „groźne" dla nauki i nic nie zmienia, a ogólnie panująca teoria dalej funkcjonuje. W związku z tym badacz może pozwolić sobie na większy margines błędu II rodzaju. Za bardziej groźne uważa się nieuzasadnione przyjęcie H_1, bowiem może to prowadzić do podważenia ogólnie panującej teorii i wprowadzenia innych badaczy w błąd. W tym przypadku akceptowalne ryzyko pomyłki popełnienia błędu I rodzaju ustalone jest na znacznie niższym poziomie wyznaczonym przez α. Szerszy opis tego problemu, wraz z obrazowymi przykładami, można znaleźć w wielu podręcznikach statystyki, np. Stanisz (2006).

Porównania wielokrotne

W większości przypadków badacz przeprowadzający eksperymenty naukowe dąży do uzyskania takich wyników, które pozwolą na odrzucenie H_0, umożliwią ogłoszenie światu

nowego odkrycia i popchną daną dziedzinę nauki do przodu na inne, nieznane tory. Jak wspomniano powyżej, maksymalne ryzyko pomyłki, na jakie zwykle godzimy się, przyjmując H_1, wynosi $\alpha = 0,05$. Wynika z tego, że „prawdopodobieństwo nie pomylenia się" jest równe $p = 1 - 0,05 = 0,95$. Należy jednak pamiętać, że wartości podane powyżej określone są w sytuacji, gdy przeprowadzany jest jeden i tylko jeden test statystyczny. Zgodnie z podstawowymi zasadami rachunku prawdopodobieństwa, w przypadku dwukrotnego testowania hipotez dochodzi do koniunkcji zdarzeń. Zakładając sytuację idealną, w której wynik jednego testu nie wpływa na wynik drugiego testu (testy są niezależne), to prawdopodobieństwo niepomylenia się ani raz przy dwukrotnym testowaniu statystycznym wynosi $p = (1- 0,05) \times (1 – 0,05) = 0,95^2 = 0,9025$. Łatwo obliczyć, że w takim przypadku prawdopodobieństwo pomylenia się przynajmniej raz jest równe $p = 1- 0,95^2 = 1- 0,9025 = 0,0975$. Dla większej liczby testów wielokrotnych prawdopodobieństwo pomylenia się, przynajmniej raz na k porównań, szybko rośnie wraz z liczbą przeprowadzanych testów, według ogólnego wzoru $p = 1-(1- \alpha)^k$. Przykładowo: dla (zaledwie) $k = 6$ porównań i poziomu istotności $\alpha = 0,05$, ryzyko błędnego odrzucenia hipotezy zerowej i przyjęcia co najmniej jednej fałszywej hipotezy alternatywnej (tj. pojawiania się wyniku istotnego statystycznie wyłącznie przez przypadek) wynosi aż $p = 0,265$. Jest to wartość znacznie wyższa niż przyjęty margines błędu I rodzaju (Stanisz, 2007).

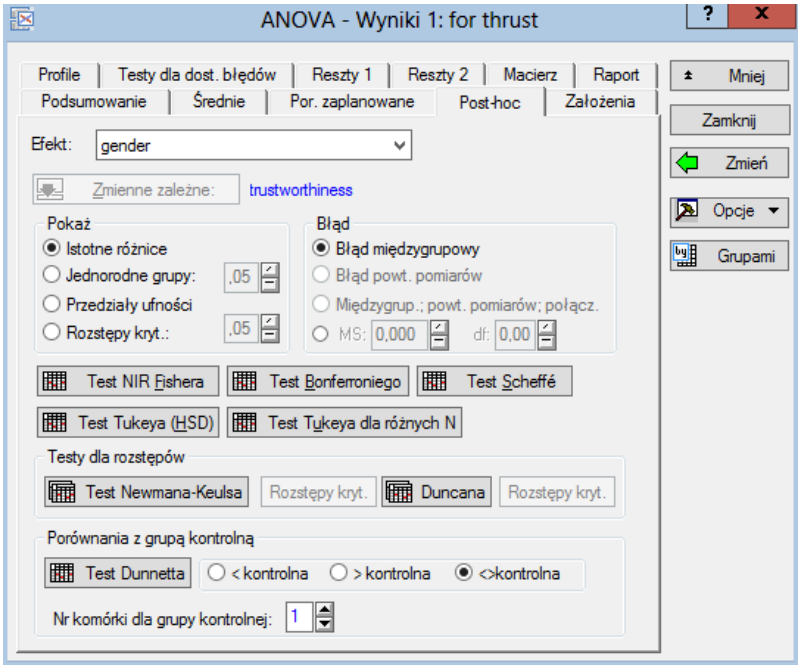
W tym miejscu warto zadać sobie pytanie, kiedy badacz spotyka się z problemem porównań wielokrotnych? Parafrazując Tukeya (1977), można powiedzieć, że problem ten pojawia się za każdym razem, kiedy na podstawie danych z jednego eksperymentu przeprowadzamy więcej niż jeden test statystyczny. Przykładowo: jeśli badacz przeprowadza badania krwi u 1000 kobiet i 1000 mężczyzn, które mają dać odpowiedź na pytanie, czy istnieją istotne statystycznie różnice międzypłciowe w liczbie erytrocytów na daną jednostkę objętości krwi, i jednocześnie przeprowadza testy statystyczne porównujące liczbę leukocytów w obu płciach, to mamy do czynienia z porównaniem wielokrotnym. Oczywiście dodanie innych porównań międzypłciowych, np.: różnic w liczbie limfocytów, trombocytów, etc., automatycznie zwiększa liczbę porównań i prowadzi do coraz większego ryzyka popełnienia co najmniej jednego błędu I rodzaju. Analogiczny problem pojawi

się przy analizowaniu korelacji pomiędzy zmiennymi. Na przykład, korelując wysokość ciała z 6 cechami temperamentu określonymi kwestionariuszowo, badacz wykonuje 6 porównań wielokrotnych, a prawdopodobieństwo otrzymania co najmniej jednego wyniku istotnego statystycznie, wyłącznie przez przypadek, jest równe $p = 0{,}265$.

Metody minimalizacji błędu I rodzaju w porównaniach wielokrotnych

Przytoczone powyżej, a także definicja porównań wielokrotnych pokazują, że w badaniach biomedycznych, jak również w badaniach z innych dziedzin, problem seryjnego testowania hipotez z użyciem jednego zestawu danych może być, a w zasadzie jest, zjawiskiem powszechnym. W związku z tym funkcjonuje wiele statystycznych metod kontroli ryzyka popełniania błędu I rodzaju.

Sztandarowymi przykładami badań, w których występuje problemem porównań wielokrotnych, są eksperymenty czynnikowe, takie jak np. ANOVA. W modelach tego typu porównania wielokrotne pomiędzy średnimi przeprowadzane są niejako „z automatu" w ramach analizy *post-hoc* (rys. 1), wykonywanej po uzyskaniu istotnego statystycznie wyniku testu globalnego *F*. Istnieje wiele metod analizy *post-hoc*, które pozwalają kontrolować ryzyko popełnienia błędu I rodzaju. Pakiet Statistica oferuje spory zestaw najczęściej stosowanych procedur. Szczegółowy opis i porównanie poszczególnych technik można znaleźć w innych publikacjach, np. Stanisz (2007).



Rys. 1. Metody porównań wielokrotnych w ramach analizy *post hoc* dla modeli ANOVA.

Niestety charakter przeprowadzanego eksperymentu lub typ uzyskiwanych danych nie zawsze dają możliwość stosowania modeli typu ANOVA i zintegrowanych z nimi technik porównań „po fakcie". Często badania polegają na wielokrotnym testowaniu danych pochodzących z jednego eksperymentu z użyciem względnie prostych metod statystycznych i bez formalnego budowania modelu statystycznego. Przykładowo: badacza może interesować wykonanie kilku testów chi-kwadrat lub przetestowanie istotności statystycznej wielu korelacji pomiędzy zmiennymi zebranymi w ramach pojedynczego badania. W takich przypadkach również stosuje się rożne metody kontroli błędu I rodzaju. Poniżej zostaną szerzej zaprezentowane dwie z nich, tj.: poprawka Bonferroniego oraz sekwencyjna poprawka Bonferroniego nazywana też metodą Holma-Bonferroniego (Holm, 1979). Są one stosunkowo łatwe do samodzielnego zastosowania i cieszą się one dużą popularnością wśród badaczy.

Poprawka Bonferroniego

W porównaniach wielokrotnych klasyczna poprawka Bonferroniego jest w zasadzie najprostszym sposobem kontroli błędu I rodzaju. Procedura jest bardzo prosta i polega na zredukowaniu wartości α poprzez podzielenie jej przez liczbę wykonywanych testów wielokrotnych k (liczbę testowanych hipotez) (Bland & Altman, 1995). Na przykład kiedy badacz korzysta z tych samych danych i chce ustalić istotność statystyczną $k = 5$ współczynników korelacji Pearsona, to istotnymi statystycznie wynikami *na poziomie istotności* $\alpha = 0{,}05$ będą jedynie takie wyniki, które pojawią się z prawdopodobieństwem mniejszym niż $p < \alpha/k = 0{,}05/5 = 0{,}01$.

Poprawka Holma-Bonferroniego (sekwencyjna poprawka Bonferroniego)

Logika redukcji błędu I rodzaju, poprzez podzielenie wartości α przez liczbę przeprowadzanych porównań wielokrotnych k, jest również wykorzystywana w przypadku procedury Holma-Bonferroniego (Holm, 1979). Jednakże w tym przypadku redukcja poziomu alfa następuje indywidualnie dla każdego testu. Dzieje się to w sposób sekwencyjny. W pierwszym kroku procedury ustalany jest poziom istotności statystycznej α. Następnie wartości prawdopodobieństwa testowego p_k, uzyskane dla kolejnych k testów (porównań wielokrotnych), szeregowane są w kolejności rosnącej (od wartości najmniejszej do największej). W pierwszej kolejności rozpatrywana jest najniższa wartość p_1. Jeżeli i tylko jeżeli wartość p_1 jest mniejsza od poziomu α podzielonego przez liczbę testów k (czyli $p_1<\alpha/k$), należy uznać, że wynik uzyskany w teście 1 jest istotny statystycznie *na założonym poziomie* α. Jeżeli warunek $p_1 < \alpha/k$ nie jest spełniony, procedura jest przerywana, a wszystkie testy uznawane są za nieistotne statystycznie. W przeciwnym wypadku procedura jest kontynuowana dla drugiego w kolejności testu o odpowiednio najniższym prawdopodobieństwie testowym p_2. Jednak w tym kroku wartość α dzielona jest przez wartość k pomniejszoną o 1, tj. (k-1), czyli sprawdzana jest nierówność $p_2 < \alpha/(k\text{-}1)$. Jeżeli nierówność nie jest spełniona, wynik testu 2 oraz wyniki pozostałych testów uznawane są za nieistotne statystycznie – procedura jest przerywana. W przeciwnym wypadku test 2 jest uznawany za statystycznie istotny, a procedura kontynuowana aż do momentu, w którym warunek p_i

$< \alpha/(1 + k\text{-}i)$ nie jest spełniony (Rice, 1989). Przykładowo: dla danych z tabeli 1, ułożone rosnąco wartości p dla $k = 5$ porównań wielokrotnych, tworzą ciąg: $p_3 < p_4 < p_1 < p_2 < p_5$. Zredukowany poziom istotności statystycznej dla najniższej w ciągu wartości ($p_3 = 0{,}003$) wynosi $\alpha/k = 0{,}05/5=0{,}01$. Ponieważ $p_3 < \alpha/k$, to wynik testu 3 jest istotny statystycznie, a procedura kontynuowana. Następna w kolejności wartość $p_4 = 0{,}011$ porównywana jest z poziomem istotności zredukowanym do wartości $\alpha/(k\text{-}1) = 0{,}05/4 = 0{,}0125$. Również w tym przypadku $p_4 < \alpha/(k\text{-}1)$, zatem wynik Testu 4 należy uznać za istotny, a procedurę kontynuować. W kolejnym kroku analizowana jest wartość $p_1 = 0{,}041$, którą porównuje się do wartości $\alpha/(k\text{-}2) = 0{,}05/3=0{,}0167$. W związku z tym, że wartość p_1 jest większa od wartości $\alpha/(k\text{-}2)$, wyniku testu 1 nie można uznać za istotny statystycznie. Procedura jest przerywana, a wyniki pozostałych testów (testu 2 oraz testu 5) uznawane są za nieistotne statystycznie.

Tabela 1. Porównanie efektów zastosowania klasycznej oraz sekwencyjnej poprawki Bonferroniego. Gwiazdka (*) oznacza, że wynik testu jest istotny statystycznie; n.i. - oznacza, że wynik testu jest statystycznie nieistotny.

Liczba porównań wielokrotnych $k = 5$	Prawdopodobieństwo testowe p_k obliczone dla poszczególnych testów	Klasyczna poprawka Bonferroniego		Sekwencyjna poprawka Bonferroniego	
		Zredukowany poziom $\alpha = 0{,}05$	Wynik	Zredukowany poziom $\alpha = 0{,}05$	Wynik
Test 1	$p_1 = 0{,}041$	$\alpha = 0{,}01$	n.i.	0,05/3=0,0167	n.i.
Test 2	$p_2 = 0{,}156$	$\alpha = 0{,}01$	n.i.	---	n.i.
Test 3	$p_3 = 0{,}006$	$\alpha = 0{,}01$	*	0,05/5=0,01	*
Test 4	$p_4 = 0{,}011$	$\alpha = 0{,}01$	n.i.	0,05/4=0,0125	*
Test 5	$p_5 = 0{,}344$	$\alpha = 0{,}01$	n.i.	---	n.i.

Problem „subiektywnej istotności statystycznej” przy porównaniach wielokrotnych

Porównanie rezultatów uzyskiwanych po zastosowaniu klasycznej i sekwencyjnej poprawki Bonferroniego zwraca uwagę na podstawowy problem występujący w przypadku metod kontroli błędu I rodzaju. Wynika on z konieczności redukcji poziomu α w zależności od liczby rozpatrywanych porównań wielokrotnych. Analizując tabelę 1, łatwo zauważyć, że w wyniku zastosowania jednej bądź też drugiej procedury można podjąć różne decyzje o istotności statystycznej uzyskanego wyniku. W przypadku pierwszej metody, zaledwie test 3 został uznany za istotny statystycznie, natomiast w metodzie sekwencyjnej statystycznie istotny był również test 4. Mówiąc językiem statystyki, klasyczna poprawka Bonferroniego jest bardziej konserwatywna (tj. trudniej jest odrzucić H_0) niż metoda sekwencyjna. Na pierwszy rzut oka może wydawać się, że owe zwiększenie konserwatywności jest właśnie zjawiskiem pożądanym. Należy jednak pamiętać, że z redukcją poziomu α nierozłącznie wiąże się zwiększenie ryzyka popełnienia błędu II rodzaju, a więc nieodrzucenia błędnej H_0 i przeoczenia rzeczywiście istniejących różnic czy zależności.

Łatwo również zauważyć, że w rozpatrywanych przypadkach „konserwatywność” silnie zależy od liczby przeprowadzanych porównań wielokrotnych. Dla stosunkowo niewielkiego zwiększenia liczby porównań wielokrotnych (np. do $k = 10$) progowy, skorygowany poziom istotności statystycznej dla obu metod wynosiłby $\alpha/k = 0{,}05/10 = 0{,}005$. Taka granica istotności dla wielu przeprowadzanych eksperymentów może być bardzo trudna do przekroczenia. Można to sobie łatwo uzmysłowić, dopisując do danych w tabeli 1 kolejne 5 testów o wartościach p z przedziału $0{,}005 < p_i < 0{,}009$. Mimo wykrycia 8 na 10 wyników testów, których indywidualne prawdopodobieństwo pojawienia się p_i jest (znacznie) niższe od założonego nieskorygowanego poziomu istotności $\alpha = 0{,}05$, po zastosowaniu omawianych poprawek żaden z przeprowadzonych testów nie mógłby zostać uznany za istotny statystycznie.

Powyższy problem może rodzić bardzo poważne konsekwencje natury filozoficzno-etycznej i finansowej. Stały się one podstawą krytyki technik redukcji poziomu α w zależności

od liczby porównań wielokrotnych (Moran, 2003; Nakagawa, 2004; Perneger, 1998). Łatwo sobie wyobrazić sytuację, kiedy początkowo istotne statystycznie korelacje, np. pomiędzy wysokością ciała a 6 cechami temperamentu, staną się statystycznie nieistotne tylko dlatego, że jednocześnie badacz postanowi przeanalizować w tym samym badaniu 5 cech osobowości. W naukach medycznych konsekwencje bezrefleksyjnego redukowania poziomu α mogą być jeszcze poważniejsze i decydować o zdrowiu i życiu pacjentów. Na przykład: stwierdzenie statystycznie istotnej większej skuteczności metody A nad metodą B może zależeć jedynie od tego, ile metod leczenia jednocześnie badacz postanowił uwzględnić w analizie danych. Analogicznie dziesiątki milionów złotych wydane na badania nad nowym lekiem mogą zostać zmarnowane, ponieważ ostateczne wnioski o jego większej skuteczności w porównaniu do innego leku lub placebo będą zależały tylko od tego, ile specyfików włączono do porównań. W obu przypadkach faktyczny efekt leczniczy może zostać przeoczony i poprzez popełnienie błędu II rodzaju szansa na pomoc i ulgę w cierpieniach wielu osób zostanie zaprzepaszczona.

W podobnym duchu można rozważać kwestie uczciwości naukowej i moralności samego naukowca (Moran, 2003; Nakagawa, 2004; Perneger, 1998). Tutaj problem opiera się na paradoksie polegającym na tym, że badacz przeprowadzający szczegółowo i rzetelnie swoje badania oraz wykorzystujący z maksymalną efektywnością zebrane dane, a więc badacz oszczędzający publiczne czy też prywatne środki finansowe, jest „karany" za swoją dodatkową pracę i uczciwość statystyczną. Przykładowo: badacz, mając do dyspozycji dane, które umożliwiają obliczenie 4 współczynników korelacji po przeprowadzeniu testów statystycznych, otrzymuje wartości p wskazujące na istotność statystyczną 2 z 4 badanych współzależności (np. $p_1 = 0{,}03$; $p_2 = 0{,}16$; $p_3 = 0{,}26$; $p_4 = 0{,}02$). W takiej sytuacji zastosowanie poprawek na porównania wielokrotne spowoduje, że żadna z wykrytych zależności nie spadnie poniżej skorygowanego poziomu α. Badacz stanie zatem przed następującym dylematem: czy opisać tylko dwa istotne testy, czy też rzetelnie opisać (nieistotne) wyniki 4 przeprowadzonych analiz. Może również rozważyć napisanie dwóch osobnych artykułów, z których każdy opisze tylko jedną istotną statystycznie współzależność. Kolejny problem pojawi się, gdy w przyszłości na tym samym materiale zostanie

doliczony jeszcze jeden dodatkowy test współczynnika korelacji. Czy wtedy należy zweryfikować wcześniejsze wnioski, wycofać wcześniejsze artykuły lub napisać do nich erratę? Powyższe rozważania ocierają się o absurd, jednak współcześnie, gdy kariera i sukces badacza w ogromnej mierze zależą od liczby publikacji, pokusa kawałkowania danych (*data slicing*) lub łowienia i raportowania tylko istotnych statystycznie wyników (*data fishing*) i zatajania nieistotnych (*publication bias*) może być zbyt silna by się jej oprzeć.

Błąd *α* czy *β* - gdzie leży granica rozsądku?

Powyższe rozważania pokazują błędne koło, w jakim często przychodzi funkcjonować naukowcom. Chcąc uniknąć kompromitacji błędu I rodzaju i nieuzasadnionego odrzucenia prawdziwej hipotezy zerowej, skazują się na marazm niekończących się korekt i redukcji poziomu istotności. To w konsekwencji prowadzi do błędu II rodzaju i niemożności przyjęcia jakiejkolwiek hipotezy alternatywnej. Na szczęście istnieje co najmniej kilka sposobów pozwalających na wyjście z takiego impasu.

Zachować zdrowy rozsądek

Choć bezrefleksyjne stosowanie „poprawek na porównania wielokrotne" może znacząco podnieść ryzyko przeoczenia istotnych statystycznie efektów, to w pewnych sytuacjach użycie owych poprawek jest w pełni uzasadnione (Perneger, 1998). Na przykład: diagnozując zdrową osobę 40 testami, należy być świadomym, że co najmniej jeden wynik istotny statystycznie (na poziomie $\alpha = 0{,}05$) może pojawić się przez przypadek z prawdopodobieństwem $p = 0{,}87$. Zastosowanie poprawki na porównania wielokrotne pomoże powstrzymać lekarza od wdrożenia niepotrzebnej procedury leczenia, choć w przypadkach związanych ze zdrowiem pacjenta należałoby potencjalnie przypadkowy wynik powtórzyć. Podobnie w badaniach behawioralnych, wykonując baterię testów psychologicznych na określonej grupie osób i uzyskując jeden istotny wynik na kilkadziesiąt kwestionariuszy, należy być świadomym, że wynik ten może być czystym przypadkiem. Poprawki na porównania wielokrotne pozwolą nam sceptycznie podejść do tak uzyskanych wyników. Innym przy-

kładem zdroworozsądkowego wykorzystania poprawek na porównania wielokrotne jest sytuacja, kiedy badacz eksploruje dane bez postawionych uprzednio konkretnych hipotez badawczych. W takim przypadku istotne statystycznie efekty korelowania i porównywania „wszystkiego z wszystkim" powinny być skorygowane przy użyciu odpowiednich technik redukujących ryzyko popełnienia błędu I rodzaju.

Rozważyć liczbę testów istotnych statystycznie, a nie tylko ich wartość p

Przeprowadzając porównania wielokrotne, badacz może spotkać się z sytuacją, w której zamiast jednej bardzo niskiej wartości *p* uzyska większą liczbę wartości *p*, niewiele mniejszych od założonego poziomu istotności, np. $\alpha = 0{,}05$. W tym drugim przypadku zastosowanie poprawek na porównania wielokrotnie redukujących poziom α może skutkować tym, że żadna z tych wartości nie przekroczy granicznego poziomu istotności statystycznej. Przykładem mogą być wyniki badań nad związkiem uwagi i koncentracji (15 zmiennych uzyskanych z testu D2) z aktywnością układu autonomicznego, wyrażoną przez jeden z parametrów zmienności rytmu serca (HRV-LF%)[14]. Jak zaprezentowano w tabeli 2, na 15 współczynników korelacji rang Spearmana, w 7 przypadkach wyniki okazały się istotne statystycznie na poziomie $\alpha = 0{,}05$. Zastosowanie poprawek dla 15 porównań wielokrotnych obniża poziom istotności statystycznej do wartości $\alpha = 0{,}05/14 = 0{,}004$. Prosta analiza tabeli 2 pokazuje, że na takim poziomie żaden ze współczynników korelacji nie może być uznany za istotny statystycznie. Czy postępując konserwatywnie i ignorując 7 na 15 wyników istotnych statystycznie, nie popełnimy błędu II rodzaju?

[14] Wyniki pochodzą z badań statutowych przeprowadzonych w Katedrze i Zakładzie Fizjologii Uniwersytetu Medycznego im Piastów Śląskich we Wrocławiu. Za udostepnienie danych dziękuję pani dr Agnieszce Siennickiej.

Tabela 2. Wartość współczynników korelacji rang Spearmana oraz poziomy *p* dla zmienności rytmu serca (miara: HRV-LF%) oraz wartości globalnej i cząstkowych testu uwagi i koncentracji D2. Gwiazdki (*) oznaczają wyniki istotne statystycznie na poziomie α = 0,05.

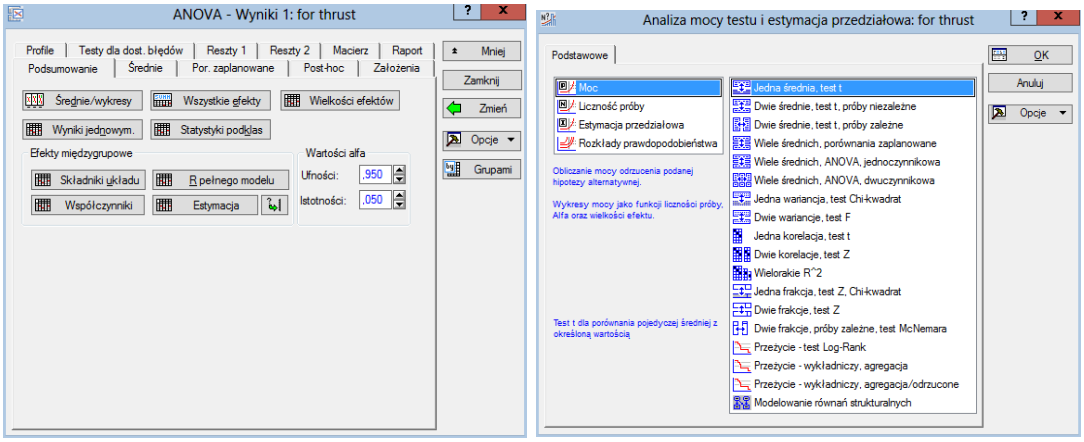
Wynik testu D2 vs. HRV-LF%	r_s	*p*
D2 - globalny	-0.34	0,016*
D2 - 1	-0.29	0,036*
D2 - 2	-0.31	0,027*
D2 - 3	-0.35	0,011*
D2 - 4	-0.30	0,033*
D2 - 5	-0.25	0,083
D2 - 6	-0.20	0,169
D2 - 7	-0.22	0,120
D2 - 8	-0.23	0,108
D2 - 9	-0.21	0,143
D2 - 10	-0.22	0,115
D2 - 11	-0.28	0,049*
D2 - 12	-0.09	0,513
D2 - 13	-0.17	0,227
D2 - 14	-0.32	0,022*

Moran (2003) podaje prosty sposób pozwalający obliczyć, jakie jest prawdopodobieństwo uzyskania określonej liczby testów istotnych statystycznie na całkowitą liczbę przeprowadzonych porównań wielokrotnych. Odnosi się on bezpośrednio do schematu Bernoulliego opisanego wzor $P_B(A) = \frac{B!}{A!(B-A)!}\,\alpha^A(1-\alpha)^{B-A}$, gdzie A oznacza liczbę „sukcesów", czyli wyników istotnych statystycznie (poniżej założonego α), a B - ogólną liczbę wykonanych testów. Warto zwrócić uwagę, że prawdopodobieństwo sukcesu jest tu definiowane

jako prawdopodobieństwo uzyskania wyniku istotnego przez czysty przypadek, czyli prawdopodobieństwo popełnienia błędu I rodzaju określonego poziomem α. Wykorzystując powyższy wzór, łatwo obliczyć, że szansa na uzyskanie A = 7 wyników istotnych na poziomie istotności α = 0,05 na B = 15 przeprowadzonych porównań wielokrotnych jest znikoma i wynosi jedynie $P_{15}(7)$ = 0,000003. Można zatem podejrzewać, że wykryte zależność odzwierciedlają rzeczywiste efekty biologiczne.

Określić wielkość stwierdzonego efektu

Przekroczenie upragnionej granicy istotności statystycznej, czyli osiągnięcie prawdopodobieństwa testowego o wartości $p < 0,05$, informuje jedynie o tym, że badane zjawisko opisane w H_1 najprawdopodobniej istnieje. Wartość *p* nie określa natomiast, jak silny i znaczący dla danej dziedziny jest zaobserwowany efekt. Wielu autorów (Garamszegi, 2006; Nakagawa, 2004; Nakagawa & Cuthill, 2007; Sullivan & Feinn, 2012) postuluje, że równie ważne albo nawet ważniejsze od raportowania wartości *p*, jest umieszczanie w pracach naukowych informacji o wielkości badanego efektu (*effect size*). Istnieje szereg technik pozwalających obliczyć wielkość obserwowanych efektów. Obliczenia mogą być przeprowadzane zarówno dla eksperymentów nakierowanych na stwierdzanie różnic pomiędzy badanymi grupami, jak i tych, które badają współzależności pomiędzy zmiennymi. Dodatkowo, cennym dla wnioskowania statystycznego będzie obliczenie przedziałów ufności dla obliczonych wielkości efektu. Przedziały te umożliwią określenie zakresów wokół wyliczonych wartości, w których z określonym prawdopodobieństwem (np. 95%) będzie znajdowała się prawdziwa (tj. ogólnopopulacyjna) wielkość efektu. Bezpośrednie porównanie przedziałów ufności obliczonych dla różnych wielkości efektów pozwala odpowiedzieć na pytanie, czy określone wielkości efektów różnią się od siebie istotnie statystycznie. W takiej sytuacji przedziały ufności na siebie nie zachodzą. Bardziej szczegółowy opis poszczególnych miar wielkości efektu oraz sposobów obliczania przedziałów ufności można znaleźć np. w pracach Nakagawa i Cuthill (2007), Sullivan i Feinn (2012) lub podręcznikach statystyki. Obliczenia wielkości badanych efektów oferowane są także przez pakiet Statistica (poszczególne moduły analityczne oraz moduł Analiza Mocy Testu).



Rys. 2. Obliczenie wielkości efektów możliwe jest w wielu modułach pakietu Statistica.

Oszacować liczbę fałszywie przyjętych hipotez alternatywnych (False Discovery Rate)

W wielu dziedzinach nauki, np. badania genetyczne, neuroobrazowanie, proteomika, porównania wielokrotne są praktyką powszechną i prowadzoną na bardzo dużą skalę. W eksperymentach obejmujących po kilkaset i więcej porównań stosowanie metod redukcji poziomu α, uzależnionych od liczby przeprowadzanych porównań wielokrotnych, może być całkowicie pozbawione sensu ze względu na dużą konserwatywność i znaczne ryzyko popełnienia błędu II rodzaju. W takich sytuacjach stosuje się inne podejście teoretyczne, pozwalające na rozsądną interpretację uzyskiwanych wyników. Aby zrozumieć, na czym ono polega, należy jeszcze raz uświadomić sobie, że zaprezentowane dotychczas rozwiązania problemu porównań wielokrotnych skupiają się na kontroli ryzyka otrzymania co najmniej jednego przypadku niepoprawnego odrzucenia H_0. Jednakże w badaniach, w których jednoczenie przeprowadza się setki porównań wielokrotnych, interesująca dla badacza może być sama liczba nieprawidłowo odrzuconych hipotez zerowych, czyli liczba fałszywie istotnych wyników w stosunku do ogólnej liczby wszystkich wyników, które w eksperymencie zostały uznane za istotne statystycznie. Tak

obliczony stosunek nazwany *Frakcją Fałszywych Odkryć (False Discovery Rate, FDR* [15] (Benjamini & Hochberg, 1995)), może służyć do obliczenia tzw. *poziomu-q* (*q-value*). Podczas gdy *poziom-p* oznacza prawdopodobieństwo nazwania określonego wyniku istotnym w sytuacji, gdy rzeczywiście jest on nieistotny, *poziom-q* interpretuje się jako prawdopodobieństwo, że określony wynik, który został nazwany istotnym statystycznie, jest w rzeczywistości statystycznie nieistotny (Pike, 2011). Podobnie jak w przypadku *poziomu-p,* również w przypadku *poziomu-q* badacz godzi się na pewne akceptowalne prawdopodobieństwo „fałszywego odkrycia". Jeśli ryzyko fałszywego odkrycia, obliczone dla poszczególnych porównań wielokrotnych (wartość *q*), jest odpowiednio małe, np. mniejsze niż 5%, wynik określonego testu statystycznego można uznać za istotny statystycznie. W tym miejscu należy jeszcze raz podkreślić, że „fałszywe odkrycia" (błędy I rodzaju) w procesie seryjnego testowania hipotez, są zjawiskiem naturalnym, gdyż zawsze mogą pojawić się przez czysty przypadek. Trudnością jest jednak oszacowanie ich frakcji w ogólnej liczbie wyników istotnych statystycznie. Dzięki zastosowaniu odpowiednich algorytmów obliczeniowych (więcej informacji: Pike, 2011), możliwe jest pokonanie tej trudności. Od wersji 12.5, pakiet Statistica został wzbogacony o stosowne procedury obliczeniowe w postaci funkcji Visual Basic. Pozwala to na zintegrowaną z pakietem Statistica implementację koncepcji „*Frakcji Fałszywych Odkryć*" i *poziomu-q* do procesu wnioskowania statystycznego.

Porównania wielokrotne w praktyce – przykład z dziedziny biologii behawioralnej człowieka

Poniższy przykład przedstawia jeden z możliwych sposobów postępowania w przypadku pojawienia się problemu porównań wielokrotnych. Przykład pochodzi z pracy „*Does age difference really matter? Facial markers of biological quality and age difference between husband and wife*" autorstwa Danel, Dziedzic-Danel & Kleisner (2016). W pracy

[15] Wydaje się, że jak dotąd terminy *False Discovery Rate* oraz *q-value* nie doczekały się oficjalnego polskiego tłumaczenia, dlatego dla opisywanego zjawiska proponuję tłumaczenie *Frakcja Fałszywych Odkryć* oraz *poziom-q*.

analizowana była zależność pomiędzy różnicą wieku małżonków a kilkoma twarzowymi wskaźnikami tzw. jakości biologicznej obu partnerów. W skrócie uważa się, że takie cechy morfologiczne twarzy, jak: antropometryczna przeciętność, symetria, czy też stopień rozwoju cech męskich u mężczyzn (tzw. maskulinizacja twarzy) i kobiecych u kobiet (tzw. feminizacja twarzy), świadczą o pewnych atrybutach osobnika, które są korzystne z biologicznego punktu widzenia. Przykładowo atrybutami takimi mogą być tzw. „dobre geny" i większa heterozygotyczność materiału genetycznego, dające nadzieję na lepszą odporność na choroby i lepsze zdrowie, czy też wyższe poziomy hormonów płciowych świadczących o dojrzałości płciowej i możliwości płodzenia potomstwa. W związku z tym osoby mające odpowiednią konfigurację cech twarzowych są preferowane jako potencjalni partnerzy w związkach emocjonalnych. Jednakże dobór płciowy u ludzi nie opiera się jedynie na wyborze partnerów o odpowiednich cechach morfologicznych. Spośród innych ważnych cech, np. osobowość, pozycja społeczna, zasoby materialne, istotną rolę odgrywa również wiek potencjalnego partnera. Kobiety preferują mężczyzn nieznacznie (kilka lat) starszych od siebie, natomiast odwrotny kierunek zależności jest obserwowany u mężczyzn. Uważa się, że taki charakter preferencji względem wieku partnera, także ma podłoże biologiczne. Preferencje kobiet do nieznacznie starszych mężczyzn mogą odzwierciedlać preferencje do zwykle wyższego statusu społeczno-ekonomicznego starszych mężczyzn. Z drugiej strony preferowanie przez mężczyzn nieco młodszych partnerek może być wyrazem poszukiwania partnerek o wyższej płodności i tzw. potencjale reprodukcyjnym, cech, które są ujemnie skorelowane z wiekiem kobiet. Odwołując się do nakreślonego powyżej podłoża teoretycznego, autorzy postanowili zbadać, czy istnieje związek pomiędzy twarzowymi markerami jakości biologicznej oraz różnicą wieku pomiędzy partnerami.

W badaniach wykorzystano fotografie twarzy (*en face* i profile) mężczyzn i kobiet z 49 małżeństw. Na podstawie fotografii i z użyciem technik geometrii morfometrycznej na każdej twarzy zmierzono poziom asymetrii (zdjęcie *en-face*) oraz dymorfizmu płciowego (zdjęcia *en-face* i profil) i antropometrycznej przeciętności (zdjęcia *en-face* i profil). W rezultacie, dla każdej osoby (zdjęcia twarzy) uzyskano 5 wartości opisujących wspomniane

cechy morfologii twarzy. Dodatkowo, dla każdej pary została obliczona różnica wieku między mężem i żoną.

W pierwszych etapach analiza statystyczna polegała na obliczeniu 5 współczynników korelacji rang Spearmana pomiędzy różnicą wieku małżonków a zmiennymi opisującymi morfologię twarzy. Obliczenia wykonano osobno dla mężczyzn i kobiet. Wartości *p* (tabela 3), uzyskane dla poszczególnych współczynników korelacji, wskazywały na istnienie jednej istotnie statystycznej współzależności dla mężczyzn oraz trzech dla kobiet. W małżeństwach charakteryzujących się większą różnicą wieku pomiędzy małżonkami mężczyźni cechowali się zmaskulinizowanymi profilami twarzy, a kobiety miały twarze bardziej symetryczne, z bardziej morfologicznie „przeciętnym" i, co zaskakujące, zmaskulinizowanym profilem. Jednakże „istotne" wartości *p* niewiele przekraczały założony poziom istotności statystycznej $\alpha = 0{,}05$. Zastosowanie poprawek Bonferroniego na porównania wielokrotne skutkowałoby koniecznością uznania wszystkich wyników za nieistotne statystycznie. Nie chcąc przeoczyć potencjalnie istotnych korelacji i popełnić błędu II rodzaju, w dalszym kroku analiz postanowiono określić „stabilność" uzyskanych współczynników korelacji. W tym celu dla każdej z analizowanych zależności przeprowadzono procedurę bootstrapingu[16]. W dużym skrócie i uproszczeniu polega ona na wielokrotnym (tu: 10 000 razy) losowaniu ze zwracaniem danych z analizowanej próby i każdorazowym obliczaniu (na wylosowanych danych) wartości danej statystyki (tu: współczynnika korelacji rang Spearmana). Metoda taka umożliwia wygenerowanie rozkładu danej statystyki oraz obliczenie przedziałów ufności. W omawianym przykładzie obliczone w procedurze bootstrapingu przedziały ufności potwierdziły „stabilność" istotnych statystycznie współczynników korelacji (95% przedziały ufności nie zawierały zera). Co więcej, jako że współczynniki korelacji są jednocześnie miarami siły efektu, potwierdzone zostało również biologiczne znaczenie wykrytych zależności.

[16] Choć Statistica oferuje jedynie ograniczone możliwości użycia technik Bootstrapingu, to w ramach pakietu możliwa jest bezpośrednia współpraca ze środowiskiem R i wykorzystanie dostępnych tam funkcji obliczeniowych.

Tabela 3. Współczynniki korelacji rang Spearmana dla różnicy wieku oraz twarzowych markerów jakości biologicznej.

(n = 49)	Mężowie		95% PU		Żony		95% PU	
Różnica wieku vs.	r_s	p	*DG*	*GG*	r_s	p	*DG*	*GG*
Dymorfizm płciowy (*en face*)*	-0,04	0,77	-0,32	0,24	-0,13	0,38	-0,43	0,20
Dymorfizm płciowy (profil)*	0,36	0,01	0,11	0,58	-0,33	0,02	-0,58	-0,05
Przeciętność (*en face*)	0,16	0,27	-0,15	0,44	0,03	0,86	-0,24	0,30
Przeciętność (profil)	0,05	0,71	-0,22	0,31	0,33	0,02	0,07	0,54
Asymetria	-0,22	0,14	-0,49	0,07	-0,30	0,03	-0,51	-0,05

* – maskulinizacja dla mężczyzn, feminizacja dla kobiet; PU – przedział ufności; DG, GG – dolna i górna granica przedziałów ufności.

Dodatkowym potwierdzeniem tak otrzymanych wyników było przeprowadzenie za pomocą pakietu Statistica analizy regresji metodą cząstkowych najmniejszych kwadratów. W uproszczeniu metoda ta jest połączeniem analizy składowych głównych z analizą regresji. Umożliwia ona zredukowanie liczby analizowanych zmiennych do określonej liczby komponentów, które najlepiej odzwierciedlają związek pomiędzy zmiennymi niezależnymi i zmienną zależną (lub kilkoma zmiennymi zależnymi w bardziej rozbudowanych modelach). W opisywanych badaniach zestaw 10 zmiennych zależnych (twarzowych wskaźników jakości biologicznej) został zredukowany do 1 Komponentu, który wyjaśniał zdecydowanie największą część wariancji różnicy wieku małżonków. Szczegółowa analiza powiązań poszczególnych zmiennych z wartościami Komponentu 1 w większości potwierdziła wyniki uzyskane za pomocą prostych korelacji. Jedynie związek pomiędzy przeciętnością profilu twarzy kobiet a różnicą wieku małżonków nie znalazł potwierdzenia w wynikach regresji. Dodatkowo, uzyskane wyniki sugerowały, że w przypadku mężczyzn, oprócz maskulinizacji profilu, również większa przeciętność twarzy i symetria, a w przypadku kobiet większa maskulinizacja twarzy *en-face* mogą mieć związek z różnicą wieku pomiędzy małżonkami. W tabeli 4 zaprezentowano „wkłady" poszczególnych zmiennych (ładunki) w strukturę Komponentu 1 oraz adekwatne współczynniki regresji.

Tabela 4. Struktura Komponentu 1 – ładunki oraz współczynniki regresji β dla poszczególnych zmiennych uwzględnionych w modelu.

	Mężowie		Żony	
	Ładunki	β	Ładunki	β
Dymorfizm płciowy (en face)*	0.09	-0.05	-0.34	-0.13
Dymorfizm płciowy (profil)*	0.54	0.26	-0.40	-0.29
Przeciętność (en face)	0.48	0.11	0.03	-0.03
Przeciętność (profil)	0.30	0.12	0.09	0.04
Asymetria	-0.40	-0.14	-0.32	-0.15

* – maskulinizacja dla mężczyzn, feminizacja dla kobiet.

Podsumowując powyższy przykład, warto podkreślić, że bezrefleksyjne zastosowanie konserwatywnych metod kontroli błędu I rodzaju w pierwszym etapie analizy statystycznej (po wyliczeniu serii prostych korelacji) mogło doprowadzić do zaprzestania dalszych badań. Konsekwencją tego byłoby przeoczenie (najprawdopodobniej) istotnego efektu biologicznego i popełnienie nie mniej groźnego błędu II rodzaju.

Podsumowanie

Wykonując porównania wielokrotne, stosunkowo częste w naukach eksperymentalnych, badacz naraża się na przypadkowe uzyskanie wyniku istotnego statystycznie. Konsekwencją tego może być odrzucenie prawdziwej hipotezy zerowej i popełnienie błędu I rodzaju polegającego na przyjęciu nieprawdziwej hipotezy alternatywnej. Stosowane metody kontroli ryzyka wystąpienia takiego scenariusza mogą prowadzić do praktyki obsesyjnego obniżania akceptowalnego poziomu istotności statystycznej. Skutkiem tego może być popełnienie błędu II rodzaju i przeoczenie wyników faktycznie istotnych statystycznie. Mimo że istnieje kilka metod pozwalających rozwiązać problem porównań wielokrotnych, to trudno zaproponować jedną, uniwersalną i możliwą do zastosowania w każdej sytuacji.

Najlepiej więc, w poszukiwaniu statystycznej istotności, starać się nie przekroczyć granicy zdrowego rozsądku.

Literatura

1. Benjamini, Yoav, & Hochberg, Yosef. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological), 57*(1), 289-300.
2. Bland, J. Martin, & Altman, Douglas G. (1995). Multiple significance tests: the Bonferroni method. *BMJ, 310*(6973), 170. doi: 10.1136/bmj.310.6973.170.
3. Danel, D. P., Dziedzic-Danel, A., & Kleisner, K. (2016). Does age difference really matter? Facial markers of biological quality and age difference between husband and wife. *HOMO - Journal of Comparative Human Biology, 67*(4), 337-347. doi:10.1016/j.jchb.2016.05.002.
4. Garamszegi, László Zsolt. (2006). Comparing effect sizes across variables: generalization without the need for Bonferroni correction. *Behavioral Ecology, 17*(4), 682-687. doi: 10.1093/beheco/ark005.
5. Holm, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure, *Scand. J. Statist*. 6: 65-70.
6. Moran, M. D. (2003). Arguments for rejecting the sequential Bonferroni in ecological studies. *Oikos, 100*(2), 403-405. doi: 10.1034/j.1600-0706.2003.12010.x
7. Nakagawa, Shinichi. (2004). A farewell to Bonferroni: the problems of low statistical power and publication bias. *Behavioral Ecology, 15*(6), 1044-1045. doi: 10.1093/beheco/arh107.
8. Nakagawa, Shinichi, & Cuthill, Innes C. (2007). Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biological Reviews, 82*(4), 591-605. doi: 10.1111/j.1469-185X.2007.00027.x

9. Perneger, Thomas V. (1998). What's wrong with Bonferroni adjustments. *BMJ, 316*(7139), 1236-1238. doi: 10.1136/bmj.316.7139.1236.

10. Pike, Nathan. (2011). Using false discovery rates for multiple comparisons in ecology and evolution. *Methods in Ecology and Evolution, 2*(3), 278-282. doi: 10.1111/j.2041-210X.2010.00061.

11. Rice W.R. (1989). Analyzing Tables of Statistical Tests, *Evolution*, 43 (1), 223-225.

12. Stanisz Andrzej. (2006). *Przystępny kurs statystyki z zastosowaniem Statistica PL na przykładach z medycyny. Tom 1. Statystyki podstawowe.* Kraków: Statsoft Polska Sp. z o. o.

13. Stanisz Andrzej. (2007). *Przystępny kurs statystyki z zastosowaniem Statistica PL na przykładach z medycyny. Tom 2. Modele liniowe i nieliniowe.* Kraków: Statsoft Polska Sp. z o. o.

14. Sullivan, Gail M., & Feinn, Richard. (2012). Using Effect Size or Why the P Value Is Not Enough. *Journal of Graduate Medical Education, 4*(3), 279-282. doi: 10.4300/JGME-D-12-00156.1.

15. Tukey, J. W. (1977). Some Thoughts on Clinical Trials, Especially Problems of Multiplicity. Science, 198(4318), 679-684. doi: DOI 10.1126/science.333584.