Sylwia Bedyńska
Jakub Niewiarowski
Marzena Cypryańska
Szkoła Wyższa Psychologii Społecznej

ROZDZIAŁ

1

Wprowadzenie do analizy wariancji

W tym rozdziale znajdziemy:

- wprowadzenie do idei analizy wariancji
- z czego wynika przewaga analizy wariancji w stosunku do testu *t*-Studenta
- omówienie znaczenia efektów interakcyjnych
- omówienie znaczenia mocy testu i jego determinant
- porównanie tekstów parametrycznych i nieparametrycznych.

ANALIZA WARIANCJI JAKO TECHNIKA PORZĄDKOWANIA DANYCH

Wiele istotnych wynalazków czy ważnych odkryć naukowych w ogóle nie zostałoby dostrzeżonych przez opinię publiczną, gdyby nie benedyktyńska praca uczonych, którzy zdając sobie sprawę z doniosłości odkrycia, opisywali je, promowali, pomagali wdrożyć. Podobny los spotkał też analizę wariancji – jedną z popularnych metod statystycznej analizy danych, która została stworzona przez biologa i statystyka zarazem – Ronalda A. Fishera. Jak podają liczne anegdoty, znany ze swej ekscentryczności Fisher potrafił zapisywać nawet najważniejsze kwestie na małych skrawkach papieru, które często wyrzucał później przez przypadek do kosza. Choć sam nie był człowiekiem zbyt uporządkowanym, doskonale zdawał sobie sprawę z wagi porządku czy raczej uporządkowania. Był niezmordowanym badaczem, zwolennikiem metody eksperymentalnej i uwielbiał rozwiązywać praktyczne problemy. Analiza wariancji powstała właśnie jako praktyczna metoda pozwalająca rozstrzygać kwestie wpływu rozmaitych czynników na szybkość wzrostu roślin oraz wielkość plonów. Początkowo metoda ta była opracowana jedynie w formie podręcznika służącego analizie danych w dziedzinie rolnictwa i biologii, bez wsparcia w postaci wzorów matematycznych. Fisher, by ułatwić jej zrozumienie, opisywał ją za pomocą wykresów, wykorzystując między innymi dane dotyczące wagi swojego nowo narodzonego syna. Książka była powszechnie stosowana przez biologów i osiągnęła znaczący jak na owe czasy sukces czytelniczy. Umknęła jednak uwadze statystyków matematycznych, między innymi ze względu na brak typowego sposobu przedstawiania twierdzeń matematycznych wraz z dowodami matematycznymi, uzasadniającymi poprawność tych twierdzeń. Nie wiadomo, czy Fisher zdawał sobie sprawę z wagi swojej metody, skoro sam twierdził, że analiza wariancji to jedynie metoda porządkowania danych. Technika ta została spopularyzowana dopiero przez szwedzkiego matematyka Haralda Cramera, który przedstawił metody Fishera w formie twierdzeń matematycznych, co upowszechniło analizę wariancji w kręgach matematyków statystycznych. Dlaczego ta technika została nazwana przez jej twórcę metodą porządkowania danych, pokazuje niżej umieszczony przykład.

Jak sama nazwa wskazuje, istotę analizy wariancji musi stanowić porównywanie wariancji, czyli zmienności wyników. Aby ułatwić zrozumienie idei tej analizy, zaprezentujemy ją na przykładzie porównań między grupami. Wszystkie badane przez nas zjawiska charakteryzuje jakiś poziom zmienności; ludzie mają różny wzrost, wagę, poziom inteligencji, w tej samej sytuacji jedni bywają agresywni, a inni spokojni, ale też te same osoby w jednych warunkach przypominają anioły, a w innych zachowują się jak z piekła rodem. To właśnie zmienność. Chcąc wyjaśnić jej przyczyny i uwarunkowania, próbujemy znaleźć warunki ją

porządkujące. Weźmy prosty przykład: zmienność osiągnięć szkolnych wśród uczniów gimnazjum. Możemy spodziewać się, że znajdziemy wyniki z całego wachlarza osiągnięć, od najniższych, poprzez średnie, do najwyższych. Załóżmy, że szukając odpowiedzi na pytanie, dlaczego jedni uczniowie mają niskie osiągnięcia, a inni wysokie, podzieliliśmy ich wszystkich na trzy grupy ze względu na poziom inteligencji (niski, średni, wysoki). Gdyby po wprowadzeniu tego podziału okazało się, że zróżnicowanie wyników uległo uporządkowaniu w ten sposób, że w grupie uczniów o niskim poziomie inteligencji znalazły się głównie osoby z niskimi osiągnięciami szkolnymi, średniemu poziomowi inteligencji odpowiadał w większości średni poziom osiągnięć szkolnych, a w grupie z wysokim poziomem inteligencji znalazły się osoby z wysokim poziomem osiągnięć szkolnych – moglibyśmy powiedzieć, że zmienna *poziom inteligencji* dobrze porządkuje zmienność wyników w zakresie poziomu osiągnięć szkolnych. W ten sposób bowiem okazałoby się, że wyniki w ramach wyodrębnionych grup są podobne, natomiast grupy różnią się między sobą. To pozwalałoby nam stwierdzić, że w ramach każdej grupy zmienność wyników jest mniejsza niż w całej grupie i jednocześnie zróżnicowanie wewnątrz grup jest mniejsze niż zróżnicowanie między nimi. Test *F* w analizie wariancji sprawdza właśnie stosunek wariancji wewnątrzgrupowej (zróżnicowania wyników wewnątrz grup) do międzygrupowej (zróżnicowania między grupami). Idea ta, nakreślona tutaj jedynie zdawkowo, jest szerzej opisana w rozdziale 2, w którym omówione zostały teoretyczne podstawy analizy wariancji w jej najprostszej formie – jednoczynnikowej analizie wariancji.

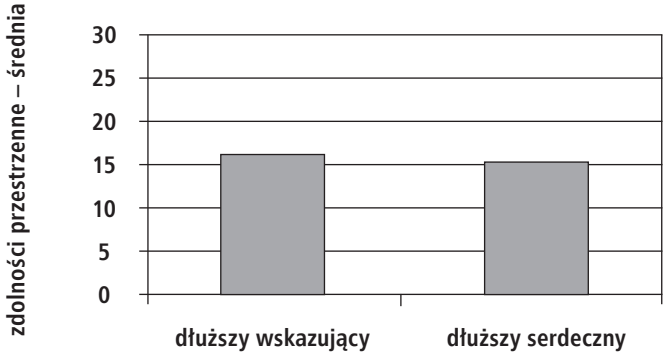
PRZEWAGA ANALIZY WARIANCJI WOBEC TESTU *T*-STUDENTA

Testy *t*-Studenta – zasłużona technika statystyczna, stosowana nie tylko do analizy znaczenia proporcji składników w procesie warzenia piwa na jego smak, ma wiele zalet. Główną jest prostota obliczeń i interpretacji w stosunku do bardziej złożonych testów statystycznych, takich jak przedstawiana w tym tomie analiza wariancji. Dodatkowo testy *t* są odporne na złamanie założeń, a także można je stosować w przypadku małych prób. Jednak nie wszystko da się sensownie analizować za pomocą porównania ze sobą pary średnich, choć cyfra 2 należy do ulubionych przez matkę naturę. Z kolei analiza wariancji daje możliwość zwiększenia liczby jednocześnie porównywanych średnich, a także uwzględnienia dodatkowych czynników. Właśnie te jej właściwości spowodowały, że analiza wariancji zdominowała testy *t*-Studenta i stała się techniką nadrzędną w stosunku do porównań parami. Poniżej opisujemy te dwa aspekty budujące przewagę analizy wariancji nad testami *t*-Studenta.

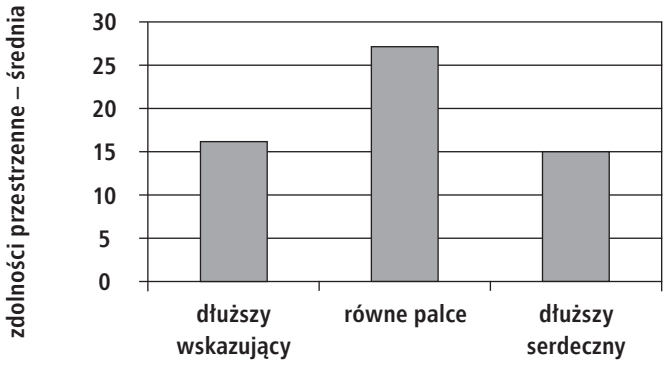
KONIEC DUALIZMU

Aby wykazać wyższość analizy wariancji nad prostymi testami *t*-Studenta, zajmiemy się problemem badawczym dotyczącym proporcji palca wskazującego i serdecznego. Ten jakże fizyczny aspekt naszego ciała ma istotne znaczenie psychologiczne, ponieważ bywa traktowany jako pośrednia miara poziomu wydzielania testosteronu w życiu płodowym dziecka. Zakłada się, że wysoki poziom testosteronu jednocześnie wpływa na taki rozwój dłoni, że wydłuża się palec serdeczny w stosunku do wskazującego i jednocześnie mózg rozwija się w taki sposób, że osoby te mają wyższy poziom zdolności przestrzennych i matematycznych w stosunku do werbalnych. Choć badania te wyglądają dość niepoważnie, a badacze biegają na badania z suwmiarkami, to jednak opublikowano sporą liczbę doniesień na ten temat w szanujących się czasopismach naukowych. Ten bardzo pośredni pomiar poziomu testosteronu w życiu płodowym dziecka jest stosowany z konieczności – bardziej inwazyjne badania hormonalne w czasie ciąży mogą być niebezpieczne dla dziecka, więc badacze raczej ich unikają.

Załóżmy, że wybraliśmy grupę osób badanych, które mają dłuższy palec serdeczny niż wskazujący, i taką, która ma odwrotną proporcję długości palców. Mierzymy poziom zdolności przestrzennych u wszystkich osób badanych na przykład testem rotacji figur, dlatego że się uważa, iż ten test w największym stopniu mierzy biologicznie uwarunkowany typ zdolności przestrzennych. Analizujemy wyniki dla dwóch grup w odniesieniu do ilościowej zmiennej określającej liczbę poprawnych odpowiedzi w teście figur. Analizujemy wyniki testem *t*-Studenta – i ku naszej rozpaczy nie uzyskujemy istotnych różnic między tymi grupami (zob. rys. 1.1). Średnie dla obu grup mają na tyle podobną wielkość, że nie możemy ich uznać za istotnie różne.



Rysunek 1.1. Hipotetyczny układ średnich dla osób o dłuższym palcu wskazującym i dłuższym palcu serdecznym



Rysunek 1.2. Hipotetyczny układ średnich dla osób o dłuższym palcu wskazującym, równych palcach i dłuższym palcu serdecznym

Zanim jednak popadniemy w czarną rozpacz, spróbujmy się chwilę zastanowić nad wykonanym właśnie porównaniem. Wzięliśmy skrajne grupy, zakładając, że wówczas będzie łatwiej znaleźć istotną różnicę. Co się dzieje jednak, w sytuacji gdy najlepszy jest pewien optymalny poziom proporcji palców? W końcu natura nie przepada za skrajnościami. Zgodnie z regresją do średniej wyniki będą wokół niej oscylowały. Zjawisko regresji do średniej, odkryte przez Francisa Galtona, wskazuje na przykład, że synowie wysokich ojców będą troszkę niżsi niż ojcowie, a synowie niskich ojców trochę wyżsi. Podobne zjawisko obserwuje się w przypadku liczby dzieci – kobiety pochodzące z wielodzietnych rodzin będą miały mniej potomstwa, a te będące jedynaczkami chętniej urodzą ich więcej. Może jest zatem tak, że skrajne grupy będą się znacząco różniły od grupy osób o podobnej długości palców wskazującego i serdecznego (zob. rys. 1.2). W takiej sytuacji konieczne jest zastosowanie jednoczynnikowej analizy wariancji, która pokaże, że środkowa grupa różni się od grup skrajnych. Jej uwzględnienie w schemacie porównań pozwoli wykazać istotne różnice. Nie zawsze więc analiza skrajnych grup się opłaci.

JEDEN CZYNNIK CZY WIELE: UROK INTERAKCJI

Wróćmy na chwilę do przykładu osiągnięć szkolnych. Na pytanie, od czego zależy poziom osiągnięć szkolnych, studenci zwykle dość szybko wskazują zmienną *poziom inteligencji*, jednakże szybko dodają: „ale też…". I tu pojawia się lista innych domniemanych czynników, które mogą być odpowiedzialne za zróżnicowanie w poziomie osiągnięć szkolnych. Świadomość owej złożoności zjawisk jest powszechna nie tylko w rozumowaniu naukowym, ale i w myśleniu potocznym. Dążąc do jak najlepszego uporządkowania obserwowanej zmienności zjawisk, rzadko poprzestajemy na prostych jednoczynnikowych schematach badawczych.

Często więc uwzględniamy w nich dwie, trzy, a nawet więcej zmiennych (choć ich nadmierne pomnażanie wcale nie jest wskazane; zob. Bedyńska i Cypryańska, 2013, rozdział 1), z których każda może być mierzona na dwóch lub kilku poziomach. W przypadkach takich wieloczynnikowych schematów badawczych do analizy możemy wykorzystać bardziej złożony typ analizy wariancji – wieloczynnikową analizę wariancji, która podobnie jak jej odmiana jednoczynnikowa może być stosowana w wersji dla zmiennych międzyobiektowych (zob. rozdział 3), w wersji dla zmiennych wewnątrzobiektowych (zob. rozdział 6), a także w tzw. schematach mieszanych, w których występuje przynajmniej jedna zmienna międzyobiektowa i przynajmniej jedna zmienna wewnątrzobiektowa (zob. rozdział 7).

WIELOCZYNNIKOWA ANALIZA WARIANCJI

W rozdziale 1 *Statystycznego drogowskazu 1* (Bedyńska i Cypryańska, 2013) opisany został przykład bardzo prostego badania dotyczącego facylitacji społecznej, czyli wpływu fizycznej obecności publiczności na poziom wykonania różnych zadań (Zajonc, Heingartner i Herman, 1969). Angażując do badania karaluchy, sprawdzano, czy obecność innych osobników wpływa korzystnie na szybkość wykonania prostych zadań.

Zadanie polegało na tym, że dany karaluch musiał przebiec specjalnie przygotowany tunel od punktu startu, gdzie pojawiało się awersyjne dlań źródło światła, do punktu końcowego, który był zaciemniony (a więc przyjazny karaluchom). Sprawdzano, czy szybkość przebycia tej trasy zwiększy się przy obecności innych karaluchów. W warunkach kontrolnych karaluch przemierzał całą trasę samotnie, w warunkach eksperymentalnych natomiast biegł w obecności „tłumu" karaluchów, umieszczonych przy trasie ucieczki w specjalnie przygotowanych boksach z przezroczystego materiału. Okazało się, że średnia prędkość, z jaką pojedynczy karaluch przemierzał trasę w obecności innych karaluchów, była istotnie większa niż w warunkach kontrolnych. Wniosek: obecność innych osobników zwiększa szybkość wykonania zadań. Szybko jednak pojawia się pytanie o uniwersalność tej reguły, czyli jakie czynniki mogą modyfikować ujawnioną zależność między szybkością wykonania zadania i obecnością innych. To jeden z najbardziej charakterystycznych aspektów postępowania badawczego: odkrywamy pewne zależności, prawa, reguły, opisujemy je i jednocześnie sprawdzamy, w jakich zakresach są one prawdziwe, jakim ograniczeniom podlegają i jak przejawia się dana zależność w zróżnicowanych warunkach wyznaczanych przez inne zmienne. W ten sposób odkrywamy kolejne zasady i całą sieć wzajemnych powiązań oraz ograniczeń w prawach rządzących interesującymi nas zjawiskami. Jedną z takich wzajemnych zależności jest **interakcja zmiennych**. Pokazuje ona wzajemne współdziałanie dwóch lub kilku zmiennych. Wróćmy jednak do przywołanego wcześniej badania Zajonca i współpracowników: okazało się, że – owszem – obecność innych osobników zwiększa szybkość wykonania zadań, ale tylko prostych (Zajonc i in., 1969). Uwzględnienie w schemacie badawczym kolejnej zmiennej *trudność zadania*

INTERAKCJA ZMIENNYCH

(zadanie łatwe *versus* zadanie trudne) pokazało ograniczenia zasady „efektywność wykonywania zadań zwiększa się w obecności innych". Okazało się bowiem, że obecność widowni poprawia wykonanie zadań prostych, natomiast pogarsza poziom wykonania zadań trudnych.

Wprowadzenie kolejnej zmiennej (*trudność zadania*) przyczyniło się więc do uchwycenia ograniczeń w zakresie korzystnego wpływu innych osobników na poziom wykonania zadań. Okazało się, że ta sama zmienna może mieć zarówno korzystny, jak i niekorzystny wpływ, zależnie od trudności zadania. Przykład ten pokazuje urok interakcji, czyli wspólnego oddziaływania dwóch lub więcej zmiennych.

OGRANICZENIA ZŁOŻONYCH PLANÓW BADAWCZYCH

Podczas rozbudowywania schematu badawczego możemy uwzględnić dowolną liczbę czynników, jednak – jak większość rzeczy w naszej rzeczywistości, również i pomnażanie zmiennych wymaga umiaru. Nie jest bowiem wskazane rozbudowywanie schematów badawczych do rozmiarów, z którymi nie sposób sobie potem poradzić. Interakcja dwóch zmiennych przybiera zwykle jasny, klarowny kształt i można by rzec: jest elegancka w swej prostocie. Interakcja trzech zmiennych bywa równie elegancka, ale może być już trudniejsza do zrozumienia (poziom trudności zależy tu od stopnia komplikacji schematu, czyli liczby poziomów każdego czynnika). Jeśli mamy szczęście uzyskać klarowny wzorzec wyników, nawet interakcja czterech czy pięciu zmiennych może być relatywnie łatwa do interpretacji, zwykle jednak – ujmując to nieco metaforycznie – przy interakcji czterech zmiennych większość relacji zaczyna spowijać mgła, przy pięciu jest już gęsta, a przy sześciu nie widać już nic. Krótko mówiąc: im bardziej złożony schemat i im więcej porównywanych grup, tym trudniej znaleźć spójną interpretację uzyskanych wyników – wraz ze wzrostem komplikacji schematu badawczego wzrasta ryzyko uzyskania wyników mało spójnych bądź zależności tworzących trudny do interpretacji wzorzec. Zwykle więc zamiast jednego monstrualnego badania zaleca się raczej serię badań eksplorujących w sposób systematyczny dany problem badawczy (zob. Wojciszke, 2004). Zalecenie to związane jest też z tym, że zwiększenie w badaniu liczby zmiennych niezależnych pociąga za sobą zwykle wzrost liczby zmiennych ubocznych czy zakłócających. To zaś utrudnia znalezienie interesujących zależności bądź wręcz fałszuje ich obraz. Takie złożone badanie, w którym mierzymy bardzo dużo zmiennych, powoduje też u osób badanych trudność ze zrozumieniem, o co chodzi, i zmniejsza realizm psychologiczny. Uwaga ta dotyczy zwłaszcza eksperymentów. Problem ten został opisany bardziej szczegółowo w rozdziale 1 *Statystycznego drogowskazu 1* (Bedyńska i Cypryańska, 2012) oraz we wprowadzeniu do analizy wariancji z powtarzanym pomiarem w rozdziale 4 niniejszego tomu.

Jeszcze jedna ważna podpowiedź. Jak napisaliśmy powyżej, trudność ze stosowaniem złożonych schematów badawczych wynika także z ograniczeń w możliwościach przetwarzania oraz integracji informacji w umyśle badacza. Trudno jest połączyć wiele cząstkowych informacji o różniących się parach średnich w jedną spójną całość. Aby ułatwić ten proces, warto się zastanowić przed analizą wyników, jak powinny one wyglądać, jeśli nasze założenia teoretyczne są słuszne; jak może wyglądać wykres średnich, jeśli wyniki ułożą się idealnie po naszej myśli. Ten proces myślowy będzie stanowił znakomite przygotowanie do analizy danych i wyszukiwania wzorców.

Przy złożonych badaniach warto także pamiętać o ważnej potrzebie człowieka, jaką jest rozumienie tego, co się robi. Jeśli badamy ludzi, musimy pamiętać, że chcą oni rozumieć sytuację badawczą, w której się znaleźli. Im więcej rozmaitych zadań, manipulacji czy kwestionariuszy otrzymuje do wypełnienia w trakcie procedury badawczej osoba badana, tym trudniej będzie jej zrozumieć zadanie, które przed nią stoi. Tym samym sytuacja badania będzie sztuczna i niejasna. Ten czynnik nie pozostaje bez znaczenia dla wyników badań. Jeśli osoba badana nie będzie mogła zrozumieć sensu swojego udziału w badaniu, może czuć się niezręcznie, co może wpłynąć na uzyskane rezultaty. Pamiętajmy, że udział w badaniu wymaga od osoby badanej sporego zaangażowania, jeśli wyniki mają możliwie dokładnie mierzyć rozmaite zmienne.

W wypadku powtarzanych pomiarów problemem będą także efekty wynikające z wielokrotnego pomiaru pewnych właściwości u tej samej grupy osób badanych. Wyobraźmy sobie, że chcemy zmierzyć samoocenę osób badanych przed wykonaniem trudnego testu i po jego wykonaniu. Osoba badana próbująca pokazać się jako spójna jednostka może celowo zapamiętać swoje odpowiedzi w kwestionariuszu i wypełnić skalę w taki sposób jak za pierwszym razem. Utrudni to analizę zmiany w poziomie samooceny. Jeśli zadaniem osoby badanej będzie rozwiązywanie przez trzy godziny serii trudnych zadań poznawczych, to niskie wyniki ostatnich będą efektem zmęczenia, a nie gorszego potencjału intelektualnego badanego. Te wszystkie efekty mogą zakłócać znacząco uzyskiwane wyniki – trzeba rozważyć możliwość ich wystąpienia już w fazie planowania badania. Szczegółowy opis trudności wynikających z powtarzanego pomiaru został zaprezentowany w rozdziale 4.

MOC TESTU I JEJ DETERMINANTY

Stanisław Lem powiedział kiedyś: „Statystyka niczego nie dowodzi, czyni tylko wszystko mniej lub bardziej prawdopodobnym". Taka jest też natura poszukiwania rozstrzygnięć naukowych – nie jesteśmy niczego pewni na sto procent, a jedynie dowiadujemy się, że jakaś różnica czy zależność jest dość prawdopodobna. Trzeba bowiem pamiętać, że poziom istotności informuje nas, jakie jest

prawdopodobieństwo uzyskania statystyki testu o danej wielkości, zakładając, że hipoteza zerowa jest prawdziwa. W ten pokrętny sposób *de facto* dowiadujemy się, że uzyskaliśmy różnicę. Istotny wynik testu mówi nam, iż jest bardzo mało prawdopodobne, że porównywane wyniki pochodzą z tej samej populacji. Badane pomiary bądź grupy różnią się nieprzypadkowo. Co jednak z wynikami nieistotnymi statystycznie? Czy wskazują one jednoznacznie, że różnice pomiędzy średnimi nie istnieją? Niestety nie. Trudności z odrzuceniem hipotezy zerowej i pokazaniem różnic czy związku między zmiennymi mogą bowiem wynikać z niedokładności pomiarowych, błędnego doboru próby, niskiej mocy testu oraz innych czynników utrudniających odszukanie efektu w szumie losowej zmienności.

BŁĄD PIERWSZEGO RODZAJU
BŁĄD DRUGIEGO RODZAJU

Warto pamiętać, że wnioskowanie statystyczne jest obciążone **dwoma błędami**: **pierwszego rodzaju**/typu/stopnia (alfa) i **drugiego rodzaju**/typu/stopnia (beta). Błąd pierwszego rodzaju polega na tym, że odrzucamy hipotezę zerową i twierdzimy, że są różnice, podczas gdy tak naprawdę tych różnic nie ma. Wielkość tego błędu określa poziom istotności. Błąd drugiego rodzaju to pomyłka polegająca na tym, że nie odrzucamy hipotezy zerowej wówczas, gdy tak naprawdę jest ona fałszywa. A zatem to taki błąd, który prowadzi do niewykazania różnic ani związku między zmiennymi. Jak widać, nieistotny wynik testu statystycznego może być efektem tego właśnie błędu, a nie braku istotnych różnic.

Omówmy te błędy na przykładzie testu ciążowego do domowego użytku. Jest on jednym z najlepszych przykładów na to, że podejmowanie decyzji statystycznych nie jest czymś w rodzaju wiedzy tajemnej przeznaczonej wyłącznie dla osób, które w pocie czoła „wkuwają na pamięć" podręczniki do statystyki. Ten test może zinterpretować każdy, kto przeczyta krótką instrukcję obsługi. Jego zadaniem jest sprawdzenie hipotezy o obecności ciąży – hipoteza zerowa to brak ciąży, a alternatywna to obecność ciąży. Jeśli test jest istotny, to ciąża jest obecna; jeśli nieistotny, to ciąża jest nieobecna. Istotność pokazywana jest tu za pomocą dwóch kresek, gdzie dwie kreski oznaczają wynik istotny. Dla przykładu w IBM SPSS Statistics istotność niektórych testów oznaczana jest gwiazdką – obecność gwiazdki mówi, że test jest istotny. Przy podejmowaniu decyzji odnośnie ciąży można albo zinterpretować wynik poprawnie – stwierdzić zgodnie ze stanem faktycznym, że ciąża jest lub jej nie ma, albo zinterpretować rzeczywistość po swojemu i narazić się na błąd. Mówimy tu o dwóch możliwych błędach: 1) możemy uznać, że test jest istotny, mimo że w rzeczywistości dany efekt nie istnieje (stwierdzenie ciąży, gdy w rzeczywistości jej nie ma); 2) możemy uznać, że test jest nieistotny, mimo że w rzeczywistości dany efekt istnieje (zignorowanie faktu bycia w ciąży). W pierwszym wypadku popełnimy błąd pierwszego rodzaju, w drugim zaś – drugiego rodzaju.

Przed popełnieniem błędu pierwszego rodzaju chroni poziom istotności. W naukach społecznych zostało przyjęte, że zgadzamy się na uznanie przypadkowych wyników za nieprzypadkowe, jeśli prawdopodobieństwo ich wystąpienia

jest mniejsze bądź równe 5% ($p \leq 0{,}05$). W testach ciążowych to prawdopodobieństwo jest dużo mniejsze (zwykle $p < 0{,}01$ – ryzyko stwierdzenia ciąży, kiedy w rzeczywistości jej nie ma, jest mniejsze niż 1:100). Oferowane na rynku testy różnią się tym, jak bardzo zabezpieczają przed wykazaniem nieistniejącej ciąży, a także tym, jak łatwo ciążę wykrywają. Te dwie możliwości są wzajemnie przeciwstawne. Test, który łatwo wykrywa ciążę, nawet w przypadku jej braku, to test liberalny. Ryzyko związane z jego użyciem jest takie, że może wykazać ciążę, która nie istnieje. Z kolei ten, który rzadko pokazuje ciążę, to test konserwatywny. Ryzyko związane z jego użyciem to niewykryta ciąża. Wybierając odpowiedni test, decydujemy o tym, na który rodzaj błędu możemy być bardziej podatni. Oczywiście, w przypadku ciąży stan faktyczny objawi się niezależnie od tego, który typ testu wybierzemy, jednakże w przypadku badań stosowanych w naukach społecznych faktyczny obraz badanego zjawiska bez zastosowania dodatkowych środków może nie być łatwy do uchwycenia.

Przyjęty w danej dziedzinie nauki poziom istotności to pewien kompromis między uznawaniem przypadku za wynik oddziaływania a uznawaniem efektu oddziaływania za przypadek. W naukach społecznych uznaje się dany wynik za nieprzypadkowy przy $p \leq 0{,}05$. Jest to kryterium oparte na doświadczeniu. Dzięki tak ustalonemu prawdopodobieństwu osiąga się najlepszy stosunek błędu pierwszego i drugiego rodzaju. Czy na tym kończą się rozważania na temat tego, dlaczego czasami badania wychodzą, a czasami nie wychodzą? Niestety, tutaj dopiero się one zaczynają. Może się na przykład zdarzyć, że w jednym badaniu nie uzyskamy istotnych statystycznie zależności, ale w innym, analogicznym badaniu – zależności te okażą się istotne. O ile przyjęty w naukach społecznych poziom istotności dobrze chroni przed błędem pierwszego rodzaju (wykryciem zależności, które nie istnieją), to nie gwarantuje on, że istniejące w rzeczywistości efekty na pewno ujawnią się w przeprowadzanych analizach. Co to znaczy? W praktyce przy zastosowaniu testu statystycznego błąd pierwszego rodzaju pojawi się rzadziej niż drugiego rodzaju.

MOC TESTU

Moc testu jest powiązana z błędem drugiego rodzaju i jest odwrotnością wielkości tego błędu. Im większy błąd drugiego rodzaju, tym mniejsza jest moc. Na przykład testy o małej mocy nie są w stanie wykryć istotnych zależności ani różnic i stosując je, narażamy się na większy błąd drugiego rodzaju. Choć oczekiwany efekt istnieje, test nie jest w stanie go wykazać. Moc testu definiowana jako zdolność do wykrywania efektu, w przypadku gdy hipoteza zerowa jest fałszywa, zależy od kilku kluczowych czynników: przyjętego poziomu istotności, siły efektu, wielkości próby, ogólnego poziomu zróżnicowania wyników (ich wariancji), właściwości rozkładu i konstrukcji testu statystycznego. Wzięcie pod uwagę tych czynników pozwala na takie zaplanowanie badania, by zwiększyć szanse na uzyskanie wiarygodnych wyników.

POZIOM ISTOTNOŚCI

Pierwszy z czynników – poziom istotności – jest bardzo istotną determinantą mocy testu i wielkości błędu drugiego rodzaju. Statystycy wskazują, że im jest niższa wartość poziomu istotności (im bardziej krytycznie testujemy swoje hipotezy), tym większa jest szansa, że nie uda nam się odrzucić hipotezy zerowej, a więc pokażemy, że różnice są nieistotne statystycznie. Będziemy więc zmuszeni utrzymać hipotezę zerową, mimo że jest ona fałszywa, ze względu na przyjęty zbyt restrykcyjny poziom istotności. Redukowanie ryzyka popełnienia błędu pierwszego rodzaju automatycznie powoduje zwiększenie ryzyka drugiego rodzaju i zmniejsza moc testu (King, Minium, 2009). Choć jest to kryterium zwyczajowe, powszechnie uznawane w naukach społecznych, to wydaje się, że poziom istotności $p \leq 0{,}05$ jest dobrym kompromisem między błędem pierwszego i drugiego rodzaju, więc nie ma sensu ustalać bardziej restrykcyjnego poziomu.

SIŁA EFEKTU

Poziom istotności, choć stanowi ważną informację pozwalającą podjąć decyzje dotyczące hipotezy zerowej i alternatywnej, ma pewne wady. Jedną z nich jest silne uzależnienie poziomu istotności od stopni swobody. Na przykład w teście *t*-Studenta dla prób niezależnych, który porównuje dwie grupy, im większa jest liczba osób badanych (większa liczba stopni swobody), tym niższy jest poziom istotności dla tej samej wartości statystyki *t*. Oznacza to, że przy tej samej różnicy średnich im większa jest liczba osób badanych, tym łatwiej jest uzyskać istotność różnic. Aby określić, jak duże są różnice między średnimi, oblicza się dodatkowe miary tzw. siły efektu, niezależne od stopni swobody i liczby osób badanych. Także istotne efekty mogą się między sobą bardzo różnić tym, na ile są silne. Silny efekt pozwala w znacznym stopniu przewidywać wyniki zmiennej zależnej, ponieważ wyjaśnia duży procent jej zmienności, słaby efekt – w znacznie mniejszym stopniu. Dodatkowo, zdecydowanie łatwiej jest empirycznie wykazać istnienie silnych efektów, natomiast słabe wymagają bardziej dokładnych narzędzi pomiarowych, większej liczby badanych oraz wykorzystania testów statystycznych o dużej mocy. Analogią mogą być poszukiwania fizyków dotyczące budowy atomu. Potrzeba było wielkiego zderzacza hadronów, by eksperymentalnie udowodnić istnienie bozonu Higgsa, dlatego że cząstka pojawia się na ułamek czasu i tylko w bardzo specyficznych warunkach, więc jej zarejestrowanie nie byłoby możliwe z użyciem mniej dokładnych narzędzi. Podobnie w psychologii niektóre efekty są tak słabe, że wymagają specjalnego podejścia, narzędzi pomiarowych czy testów statystycznych.

SIŁA EFEKTU

Załóżmy, że interesuje nas wielkość zarobków, która zależy od poziomu wykształcenia. Analizujemy różne czynniki wpływające na zarobki i widzimy, że na przykład

wykształcenie wyjaśnia 30% zmienności zarobków, podczas gdy posiadanie koneksji rodzinnych tylko 5%. Możemy wtedy porównać siłę obu czynników w przewidywaniu zarobków i dostosować swoją strategię maksymalizacji dochodów poprzez dalsze kształcenie, a nie chodzenie do cioci na imieniny. Tak więc poziom istotności pozwala podjąć decyzję zero-jedynkową: „czynnik ma znaczenie lub czynnik nie ma znaczenia", a miary siły efektu pozwalają różnicować istotne czynniki na te silniejsze i słabsze. Miary siły efektu pozwalają też zorientować się, czy udało nam się znaleźć kluczowy czynnik do wyjaśniania danego zjawiska. Jeśli siła efektów analizowanych przez nas czynników jest niewielka, to możemy spodziewać się, że wielkie odkrycie tego najważniejszego czynnika jest wciąż jeszcze przed nami.

Siła efektu ma też znaczenie przy planowaniu dalszych badań. Silne efekty okazują się istotne nawet wówczas, gdy próba jest stosunkowo niewielka, słabe efekty wymagają większej liczebności próby, by okazały się istotne. Na przykład przy słabym efekcie może się okazać, że musimy przebadać kilkaset osób, aby go ujawnić; z kolei przy analizie silnego efektu wystarczy zbadanie kilkunastu osób (Stevens, 2002). Zanim przeprowadzimy badanie, warto więc zastanowić się (np. na podstawie przeglądu dotychczasowych badań z danej dziedziny), jakiego typu efektów należy się spodziewać. W praktyce bowiem może się okazać, że brak istotnych statystycznie efektów jest konsekwencją zbyt małej próby lub niskiej siły badanego efektu.

ZRÓŻNICOWANIE WYNIKÓW

Duża wielkość zróżnicowania (wariancja) wyników w badanych populacjach jest zjawiskiem utrudniającym wykazanie istotnych efektów. Im większe jest zróżnicowanie wyników w badanych grupach, tym mniejsza okazuje się wartość statystyk testu oraz niższy poziom istotności. A zatem potrzebny jest wówczas test o dużej mocy, by wykazać istotne statystycznie efekty. Oczywiście, siła efektu i wielkość próby będą działały przeciwstawnie do wielkości wariancji wyników – będą minimalizowały wpływ wariancji, aczkolwiek przy bardzo heterogenicznych (zróżnicowanych) próbach oczekiwane różnice mogą się nie ujawnić, nawet jeśli badamy silne efekty. Warto tutaj podkreślić, że heterogeniczność badanych grup jest zjawiskiem, które bardziej wpływa na błąd drugiego rodzaju (niedostrzeganie efektu) niż błąd pierwszego rodzaju (uznanie przypadku za efekt).

NORMALNOŚĆ ROZKŁADU

Czynnikiem obniżającym moc testu jest też brak normalności rozkładu analizowanej zmiennej. Jeśli w badanej populacji rozkład danej cechy jest silnie skośny, testy parametryczne mogą okazać się zbyt słabe, by wykazać istnienie poszukiwanych efektów (w sytuacji gdy istnieją one naprawdę). W takim wypadku zaleca się

stosowanie technik nieparametrycznych, które nie mają założenia dotyczącego rozkładu danej cechy w populacji. Natomiast jeśli silna skośność objawia się w badanej próbie, to w celu zwiększenia mocy testu zaleca się dokonywanie analiz technikami parametrycznymi, jednakże na odpowiednio przekształconych danych. Na przykład jeśli analizujemy dane silnie prawoskośne (jak czasy reakcji), może się okazać, że odstępstwo od normalności rozkładu zmniejsza moc testu na tyle, że oczekiwane efekty nie ujawniają się w analizach. Natomiast jeśli te same dane odpowiednio przekształcimy (np. logarytmicznie albo pierwiastkowo), moc testu wzrośnie i może się okazać w konsekwencji, że oczekiwane efekty się ujawnią. Test okaże się istotny i będzie to lepsza alternatywa niż wykonywanie testów nieparametrycznych (por. Green, Salkind, 2003; Tabachnick, Fidel, 2007).

CHARAKTERYSTYKA PARAMETRYCZNYCH TESTÓW STATYSTYCZNYCH

Klasyczne testy statystyczne można podzielić na parametryczne (takie, które obliczamy dla zmiennych zależnych ilościowych) i nieparametryczne (dla jakościowych zmiennych zależnych). Generalnie, **testy nieparametryczne mają mniejszą moc** niż parametryczne, co oznacza, że częściej mogą „przeoczyć" prawdziwe efekty. Jednak testy parametryczne, w tym *t*-Studenta, dla prób niezależnych czy opisywana w tym tomie analiza wariancji wymagają spełnienia licznych założeń, na przykład dotyczących normalności rozkładu zmiennych oraz podobieństwa wariancji w obrębie porównywanych grup. Początkowo sądzono, że złamanie tych założeń prowadzi do znaczącego błędu w szacowaniu statystyk testu oraz poziomu istotności i rekomendowano wykorzystanie testów nieparametrycznych o mniejszej mocy. Na przykład w trakcie wykonywania testu *t* dla prób niezależnych możemy przestraszyć się różnic w wariancjach wewnątrzgrupowych i sięgnąć po test nieparametryczny (test *U* Manna-Whitneya). Okazuje się jednak, że decyzja taka może być zbyt pochopna. Jak pokazały symulacje z wykorzystaniem sztucznych danych o ustalonych parametrach ich rozkładów, testy **parametryczne są** z reguły **odporne** na łamanie założeń, co oznacza, że mimo ich niespełnienia statystyki tych testów nie są znacząco błędnie szacowane. Obawa przed błędem pierwszego rodzaju jest często u badaczy nadmierna, zwłaszcza że zbyt duża dbałość o spełnienie założeń może prowadzić do zwiększenia konserwatywności podejmowanych decyzji, a tym samym do częstszego popełniania błędu drugiego rodzaju (utraty mocy testu; por. Ferguson, Takane, 1999; Rencher, 1998; Stevens, 2002).

Zagadnienie mocy testu jest niezwykle ciekawe, z perspektywy zarówno statystyki, jak i metodologii. Czytelników zainteresowanych tym tematem odsyłamy do literatury przedmiotu (np. Brzeziński, 1997; Ferguson, Takane, 1999; Shaughnessy, Zechmeister, Zechmeister, 2002; Tabachnick, Fidell, 2007). Warto pamiętać

o wszystkich czynnikach, które wymieniliśmy, zarówno na etapie planowania badania, jak i analizy statystycznej. Wybór najbardziej adekwatnej metody analizy danych nie uratuje badania przeprowadzonego na zbyt małej próbie, a test o zbyt niskiej mocy nie pozwoli nam cieszyć się odkryciem. Nie należy też popadać w skrajność, chroniąc się przed błędem pierwszego rodzaju, ponieważ w konsekwencji może się okazać, że nie udało się nam wykazać istnienia prawdziwego efektu.