



O NIEWŁAŚCIWYM STOSOWANIU METOD STATYSTYCZNYCH

Andrzej Sokółowski

Akademia Ekonomiczna w Krakowie, Katedra Statystyki; StatSoft Polska Sp. z o.o.

Inspiracją do przygotowania tego opracowania była książka Phillipa I. Gooda i Jamesa W. Hardina pt. „*Common Errors in Statistics (and How to Avoid Them)*” wydana przez wydawnictwo John Wiley & Sons w 2003 roku, sama również nie wolna od błędów.

Zwrócimy uwagę na szereg nieścisłości i błędów spotykanych w stosowaniu metod statystycznych wykorzystywanych w badaniach naukowych i rozwiązywaniu problemów praktycznych. Porządek prezentacji będzie odpowiadał typowemu kursowi ze statystyki, jakkolwiek nie będziemy się tu zajmowali błędami popełnianymi przez studentów, o których można by napisać osobną książkę w konwencji humoru z zeszytów szkolnych.

Część pierwsza książki Gooda i Hardina ma prowokujące motto „*Don't think – use the computer*” (*Nie myśl – używaj komputera*). Współczesne programy komputerowe przygotowane dla potrzeb stosowania metod statystycznych pozwalają wykonywać obliczenia, przy których dawniej trzeba było spędzić wiele pracowitych godzin lub w ogóle ich nie podejmowano. Niestety jednocześnie pojawiło się niebezpieczeństwo bezmyślnego stosowania metod w sytuacji, gdy prawie wszystko daje się obliczyć.

Podstawowe pojęcia statystyczne

Statystyka to nauka o metodach badania prawidłowości występujących w zjawiskach masowych. Większość uczonych w swych poszukiwaniach stara się odkryć prawidłowości. Warto więc najpierw uświadomić sobie, dlaczego one występują. Na każde zjawisko oddziałują dwa rodzaje przyczyn: główne i uboczne. Te pierwsze wynikają z istoty zjawiska, działają w sposób trwały i ukierunkowany, jednakowo na wszystkie elementy badanej zbiorowości i one właśnie powodują powstawanie *prawidłowości*, nazywanych niekiedy *składnikiem systematycznym*. Przyczyny uboczne (czyli losowe) oddziałują różnie na poszczególne elementy zbiorowości, działają różnokierunkowo i w sposób nietrwały. One powodują *odchylenia od prawidłowości*, są źródłem *składnika losowego*. Należy koniecznie przed wykorzystywaniem metod statystycznych dobrze zrozumieć problem badawczy, poznać jego teorię i próbować zidentyfikować przyczyny główne oraz przyczyny uboczne.



Wydaje się oczywiste, że statystyk nigdy nie może pracować sam. On ma znać metody, ich uwarunkowania, założenia, sposób działania, zakres wyników, ale to wszystko to są tylko narzędzia. Bez znajomości merytorycznej strony zagadnienia można uzyskać wyniki łatwe do obalenia i wyśmiania przez fachowców z danej dziedziny. Statystyka jest nauką służebną, a ostateczna weryfikacja wyników jej metod następuje w dziedzinie, w której metody te zastosowano. Oczywiście nie mówimy tu o statystykach teoretycznych, którzy proponują nowe metody, wykorzystując dowody matematyczne lub badania symulacyjne, ale o statystykach „praktycznych”, którzy stosują metody statystyczne w różnych dziedzinach nauk empirycznych.

Praktyczne zastosowanie statystyki ma sens, jeżeli na podstawie części populacji zwanej *próbą*, wnioskujemy o *populacji*. To dwupoziomowe widzenie problemu jest niezbędne. Warto dbać o rozłączność oznaczeń (duże litery – małe litery), rozłączność pojęć (np. wartość przeciętna – średnia arytmetyczna) i precyzyjne definiowanie.

Zarówno populacja, jak i próba powinny być *jednorodne*. Większość badaczy dobrze rozumie pojęcie jednorodności, traktując je jednak raczej intuicyjnie. Precyzyjnie można definiować, że zbiorowość jest jednorodna wtedy, gdy wszystkie jej elementy pozostają pod wpływem działania tych samych przyczyn głównych. Na ogół jednorodność ocenia się merytorycznie, ale warto pamiętać, że statystyka dostarcza – w ramach metod taksonomicznych – wielu procedur umożliwiających kontrolę jednorodności, lub podział zbiorowości na jednorodne części.

Cechy statystyczne

Cechy statystyczne to właściwości jednostek statystycznych. Denerwujące jest nazywanie ich atrybutami – bo to w języku angielskim funkcjonuje nazwa *attributes*. Tradycyjnie cechy statystyczne dzielono na jakościowe i ilościowe. Formalnie tylko cechy ilościowe powinny być nazywane *zmiennymi*, ale przyjmuje się też określenie *zmiennie jakościowe*. Dla porządku warto pamiętać, że cechy ilościowe mają *wartości*, natomiast cechy jakościowe – *warianty*.

Podstawowe znaczenie dla późniejszego wyboru metod ma precyzyjne zdefiniowanie cech statystycznych oraz określenie skal pomiaru. Szeroko akceptowane jest rozróżnienie czterech skal pomiaru: nominalnej, porządkowej, przedziałowej i ilorazowej. Skala pomiaru determinuje na przykład wybór metody przy analizie współzależności zjawisk.

Trzeba pamiętać, że rangi, które są efektem pomiaru w skali porządkowej, nie pozwalają na liczenie odległości (a więc również różnic) i średnich. Przykładem łamania tej zasady jest rangowa metoda porządkowania obiektów wielocechowych (stosowana w jednym ze znanych rankingów szkół wyższych) oraz (o zgrozo)... współczynnik korelacji rangowej Spearmana.

Błąd terminologiczny, który prawdopodobnie jest nie do wyplenienia - przykładowo ze środowiska medycznego - to nazywanie cech statystycznych parametrami. Mówi się więc o takich parametrach chorego jak: wiek, poziom hemoglobiny itp. A to wszystko to są



cechy statystyczne. Parametry to niektóre charakterystyki liczbowe zmiennych losowych, które są modelami opisującymi kształtowanie się cech statystycznych w zbiorowości obiektów. Parametry w tym sensie to wartość przeciętna, wariancja, mediana itp.

Szeregi statystyczne

Za sprawą tragicznego sposobu spolszczenia MS Excela upowszechniła się wadliwa nazwa *szeregu statystycznego* jako *serii* (??). Tymczasem angielski wyraz *series* powinien być w tym kontekście zdecydowanie tłumaczony jako *szereg*. Seria w statystyce (jest nią na przykład ciąg odchyleń o tym samym znaku w teście serii) w języku angielskim nazywa się *run*.

Pewne nieporozumienia napotkać można przy budowie szeregów rozdzielczych dla ciągłych cech ilościowych. Klasycznym przykładem jest tu rozkład populacji według wieku. Wielu badaczy uparcie lansuje klasy szeregu w stylu: 0-4, 5-9, 10-14, 15-19 itd., zawierające „dziury”. Przecież *wiek* jest cechą jak najbardziej ciągłą i sposób podawania go z dokładnością do całych lat nie zmienia charakteru cechy. Przy przejściu na poprawne klasy: 0-5, 5-10, 10-15 itd. pojawiają się pytania, co robić z osobami, które mają wiek równy dokładnie granicy klasy. Niepisana umowa (wynikająca wszakże ze „wschodnioeuropejskiej” definicji dystrybuanty) powiada, że przedziały klasowe są lewostronnie domknięte, a prawostronnie otwarte. Tak właśnie budowany jest szereg rozdzielczy w programie *STATISTICA*.

Graficzna prezentacja danych statystycznych

To zagadnienie w zasadzie pomijamy odsyłając czytelnika do małej, ale znakomitej i sławnej książeczki Darrella Huffa (z zabawnymi ilustracjami Irvinga Geisa) *How To Lie With Statistics (Jak kłamać przy pomocy statystyki)*, wydanej trzykrotnie (1954, 1982 i 1993) przez wydawnictwo W.W. Norton & Company. Pokazano tam sposoby manipulowania wykresami dla wywołania błędnego wrażenia czytelnika na przykład o znaczeniu trendu.

Typowe błędy spotykane dzisiaj to brak rozróżnienia pomiędzy wykresem słupkowym (dotyczy cechy jakościowej i jego słupki są oddzielone od siebie) a histogramem (dotyczy cechy ilościowej i słupki przylegają do siebie), łączenie punktów na diagramie korelacyjnym oraz rozpoczynanie osi pionowej w wykresie przeżyć od liczby większej od zera (to świetny przykład manipulowania wrażeniem – często nieświadomego), a kończenie na liczbie większej od 1 (co, jak w znanym dowcipie o Studium Wojskowym dopuszcza, że w warunkach bojowych prawdopodobieństwo może być większe od jedności).



Losowy dobór próby

W każdym podręczniku statystyki znajdujemy na poczesnym miejscu wymóg *losowości* i *reprezentatywności* próby. Próba jest reprezentatywna, jeżeli jej struktura jest identyczna lub bardzo zbliżona do struktury zbiorowości ogólnej. Dzięki działaniu prawa wielkich liczb ta reprezentatywność „zapewni się sama”, jeżeli próba została dobrze wylosowana. Warunek dobrego losowania jest teoretycznie prosty – każdy element zbiorowości ogólnej powinien mieć takie samo prawdopodobieństwo wejścia do próby. Praktyczne zapewnienie realizacji tej zasady jest niekiedy bardzo trudne. Wielu badaczy wydaje się nie dostrzegać istnienia dyscypliny zwanej *metodą reprezentacyjną*, w ramach której opublikowano wiele podręczników.

W wielu badaniach losowanie próby powierzane jest wyspecjalizowanym instytutom badawczym, podobnie jak proces ankietowania. Na ogół zadania te wykonywane są poprawnie, choć nie zawadzi przeprowadzenie kontroli losowości próby już po jej pobraniu.

Przy losowaniu próby bardzo łatwo jest popełnić błędy prowadzące do niereprezentatywności. Przykładem mogą tu być ankiety telefoniczne i internetowe, których wyniki nie mogą być uogólniane na całe społeczeństwo, a tylko odpowiednio na posiadaczy telefonów lub osoby posiadające dostępu do Internetu. Szczegółowe problemy reprezentatywności próby są rozważane w wielu tekstach z zakresu metodologii badań społecznych, socjologicznych i psychologicznych.

Specyficzne kłopoty z losowością próby mają lekarze i ekonomiści. Czy pacjenci leczenia w naszym szpitalu na konkretną chorobę mogą być uważani za próbę losową? To bardzo często zadawane pytanie. Na ogół odpowiedź jest twierdząca. Trzeba tylko rozważyć, czy populacja zamieszkująca teren, z którego mamy pacjentów mniej więcej odpowiada tzw. ogółowi, oraz czy na tym terenie nie ma zewnętrznych czynników mogących zakłócać przeciętną zachorowalność i przebieg leczenia danego schorzenia. Odmienna od przeciętnej struktura wieku nie stanowi tu problemu, gdyż istnieje możliwość wykorzystywania tzw. standaryzacji według wieku (to samo dotyczy badań demograficznych).

O wiele większe kłopoty teoretyczne sprawiają ilościowe badania makroekonomiczne lub regionalne. W wielu badaniach ekonomicznych trudno jest zapewnić spełnienie losowości próby. Analizując dane statystyczne dotyczące województw Polski, bierzemy przecież pod uwagę wszystkie województwa, a nie ich próbę. W tej sytuacji niektórzy negują wręcz istnienie tutaj relacji „populacja – próba”. Warto więc, obok pojęć *zbiorowość*, *populacja*, wprowadzić jeszcze *mechanizm ekonomiczny* jako cel badań statystycznych i ekonomicznych. Badając kształtowanie się bezrobocia i jego czynników w Polsce, na podstawie danych wojewódzkich, w danych statystycznych mamy obecny wspomniany efekt działania przyczyn głównych (efekt systematyczny, strukturalny) oraz efekt oddziaływania przyczyn ubocznych (efekt losowy, przypadkowy, zakłócenia).

Wymagana liczebność próby – to kolejne częste pytanie zadawane statystykom. Dla udzielenia precyzyjnej odpowiedzi statystyk musi „odbić piłeczkę”, zadając własne pytania: do czego ta próba ma służyć (estymacja, testowanie), o jakim parametrze mamy wnioskować (próba do wnioskowania o strukturze musi być zazwyczaj wielokrotnie



większa niż do wnioskowania o poziomie zjawiska), jaka jest zmienność zjawiska i wreszcie jaki poziom ufności zakłada badacz. Często występuje strach przed małą próbą; obawa o negatywną ocenę recenzentów. Jednak przy bardzo kosztownych eksperymentach badawczych, krótkich szeregach czasowych lub rzadkich chorobach, alternatywą wnioskowania na podstawie małych prób jest zaniechanie analiz w ogóle. Trzeba tylko zdawać sobie sprawę z wpływu liczebności próby na wyniki wnioskowania. Przy małej próbie „trudno” jest udowodnić hipotezy badawcze, natomiast przy bardzo dużej próbie można wręcz wykazać istotność każdej różnicy. Większość statystyk testowych da się przekształcić w ten sposób, że po lewej stronie znajdzie się n , a wynik powie, ile potrzeba obserwacji, aby wykazać, że różnica, którą obserwujemy, jest istotna statystycznie.

Prawdopodobieństwo

Aksjomatyczna definicja prawdopodobieństwa sformułowana przez Andrieja Kołmogorowa powiada (w uproszczeniu), że prawdopodobieństwo to liczba z przedziału $[0,1]$ przyporządkowana każdemu zdarzeniu losowemu. W wielu dziedzinach pozamatematycznych uważa się, że prawdopodobieństwo to liczba z przedziału $[0,100]$, wyrażona w procentach, co z jednej strony jest pokłosiem częstościowej definicji prawdopodobieństwa, a z drugiej jest bardziej intuicyjne. Formalnie ta maniera jest bardzo denerwująca dla statystyków, lecz trudno jest skutecznie walczyć z wieloletnimi przyzwyczajeniami całego środowiska.

Zmienne losowe

Popularna definicja zmiennej losowej powiada, że jest to taka wielkość, która w wyniku „doświadczenia” może przyjmować różne wartości, przy czym przed doświadczeniem nie można z absolutną pewnością przewidzieć, jaka wartość właśnie się zrealizuje. Błędne rozumienie *zmiennej losowej* zasadza się na sądzie, jakoby w zmiennej losowej występowały tylko przyczyny losowe. Uważa się, że coś jest losowe, jeżeli jest „czysto losowe” – czyli wyniki gier liczbowych, rzutu kostką, monetą, karty, jakie otrzymujemy „na ręce”, wynik losowania kul z urny. Tymczasem wystarczy tylko „trochę” tej losowości, aby absolutnie pewne prognozowanie zjawiska było niemożliwe – i już mamy zmienną losową.

Niestarannością, która utrudnia lekturę wielu prac, jest niestosowanie się do raczej powszechnej konwencji, która przewiduje, że nazwy zmiennych losowych piszemy wielkimi literami (najczęściej końcowymi alfabetu), natomiast realizacje, czyli wartości zmiennych losowych – odpowiednimi literami małymi. Nieodróżnianie zmiennej losowej od jej realizacji to niestety dość częsty błąd.

Ze zmiennymi losowymi wiąże się jeszcze jeden dość częsty błąd. Parametrem położenia jest wartość przeciętna – czyli przeciętny wynik danej zmiennej losowej. W języku angielskim parametr ten nazywa się *expected value*, więc polskie tłumaczenie *wartość oczekiwana* jest jak najbardziej poprawne językowo, podobnie jak piękny polski odpowiednik – *nadzieja matematyczna*. Niestety pojęcia te prowadzą do błędnego



mniemania, jakoby wartość oczekiwana to była wartość najbardziej prawdopodobna, najczęstszy rezultat zmiennej losowej (taka wartość to *modalna*). Niejednokrotnie zaskoczenie budzi proste wyliczenie pokazujące, że wartością oczekiwaną przy pojedynczym rzucie kostką jest liczba 3,5 – jak to możliwe, skoro nigdy takiej wartości nie da się otrzymać. Tymczasem rozważanie tej liczby jako przeciętnej z niemal nieskończonej liczby rzutów jest bardziej zrozumiałe.

Estymacja

Z pewnym wahaniem stawiam kontrowersyjne pytanie – czy statystyka opisowa ma sens? Czy jest to tylko arytmetyka na zbiorach danych? Uważam, że niemal nigdy nie chodzi nam tylko o analizę tych danych, które mamy (tych 70 pacjentów, 16 województw, 130 przedsiębiorstw itp.), a tak naprawdę chcemy poznać mechanizm, który te dane wygenerował. Chcemy więc wnioskować o populacji na podstawie próby. Musimy zatem stosować metody statystyki matematycznej – estymację i weryfikację hipotez statystycznych. W estymacji konieczne jest precyzyjne odróżnianie trzech różnych (!) elementów: parametr – estymator – ocena. Tylko ten drugi jest zmienną losową i tylko jego właściwości statystyczne (zgodność, nieobciążoność, efektywność odporność) można rozważać. Licząc średnią arytmetyczną czy odchylenie standardowe w próbie, zapominamy, że zachowują się one zgodnie z precyzyjnymi prawami rachunku prawdopodobieństwa.

Jednym z najbardziej „pechowych” problemów estymacji jest szacowanie *modalnej* (czyli *wartości najczęstszej*). W sytuacji gdy próba zawiera małą liczbę wartości, z których każda jest inna, wielu statystyków twierdzi, że modalnej nie ma. Na uwagę, że próbę wylosowano z rozkładu, który ma modalną, powiadają, że próba jest zbyt mała, aby zbudować szereg rozdzielczy i zastosować znany wzór interpolacyjny. Niemal nieznanne są proste procedury umożliwiające szacowanie (a nie „wyliczenie”) modalnej z próby o dowolnej liczebności.

Testowanie hipotez statystycznych

Tu chyba spotyka się najwięcej błędów, niedoskonałości i niewłaściwego stosowania metod statystycznych.

Hipoteza statystyczna to sąd o populacji (zjawisku) sformułowany bez pełnej znajomości tej zbiorowości. Hipotezę należy sformułować przed badaniem (a jak niektórzy dobitnie podkreślają – przed włączeniem komputera). Najczęściej hipoteza badawcza wyrażona jest jako tzw. hipoteza alternatywna, a nie – jako nie pozostawiająca wyboru – hipoteza zerowa. Przed badaniem trzeba też zdecydować się, czy hipoteza alternatywna jest jednostronna (kierunkowa) czy dwustronna (bezkierunkowa). Na przykład przed policzeniem współczynnika korelacji trzeba wiedzieć, czy hipoteza badawcza brzmi: *zmienne są istotnie skorelowane*, czy też *zmienne są dodatnio skorelowane*.

Przy weryfikacji hipotez rozważa się dwa błędy. Błąd pierwszego rodzaju polega na odrzuceniu hipotezy prawdziwej. Prawdopodobieństwo jego popełnienia zakłada sam badacz,



jest ono nazywane *poziomem istotności* i oznaczane przez α . Jak wiadomo najczęściej przyjmowaną wartością jest 0,05. Błąd drugiego rodzaju polega na przyjęciu hipotezy fałszywej i oznaczany jest β . Należy z naciskiem podkreślić, że prawdopodobieństwa dotyczą błędów w procesie decyzyjnym i nie mają nic wspólnego z prawdopodobieństwami prawdziwości hipotez zerowej i alternatywnej.

Powiada się żartobliwie, że błąd czwartego rodzaju polega na zastosowaniu niewłaściwego testu. Niestety zdarza się to często. Stosuje się testy parametryczne bez sprawdzenia (lub choćby zastanowienia się nad tym) założenia o typie rozkładu wymaganego przez dany test. Niektóre testy wymagają prób o odpowiedniej liczebności i przy próbach zbyt małych odpowiednie statystyki testowe mają rozkład inny, niż się spodziewa badacz, bo tak wyczytał w podręczniku. Przy badaniu współzależności zdarza się wykorzystywanie metod niewłaściwych dla danej skali. Nieco zamieszania wprowadzają tu obiegowe, nieprecyzyjne nazwy testów – jak *test Studenta* (dla jednej wartości przeciętnej, dla dwóch średnich, istotności współczynnika korelacji, istotności współczynnika regresji – wszystkie one wykorzystują statystykę podlegającą *rozkładowi Studenta*) czy *test chi-kwadrat* (zgodności, niezależności, dla jednej wariancji, istotności zmiennej dodanej w modelu regresji itp.).

Warto tu wspomnieć o zasłyszonym od lekarzy błędzie piątego rodzaju polegającym na wyborze niewłaściwego statystyka do wykonania obliczeń do pracy doktorskiej lub habilitacyjnej.

Przy stosowaniu testów istotności można podjąć jedną z dwóch decyzji:

- ◆ *odrzuć hipotezę zerową, przyjmując hipotezę alternatywną,*
- ◆ *nie ma podstaw do odrzucenia hipotezy alternatywnej.*

Jak widać, nie jest możliwe *przyjęcie hipotezy zerowej*, a więc nie można przykładowo udowodnić równości średnich, braku korelacji czy (niestety) normalności rozkładu. Nieodrzućenie hipotezy zerowej oznacza w praktyce, że dalej nie wiemy nic konkretnego (w sensie naukowym). Ze zdziwieniem znajdujemy w wielu podręcznikach amerykańskich (głównie z zakresu statystyki dla ekonomistów) rysunki, na których część osi wartości statystyki testowej nie będąca *zbiorem krytycznym* określana jest mianem *acceptance region*, czyli przyjęcia hipotezy zerowej. Takiej decyzji w ogóle nie przewidują testy istotności i takie podejście umożliwiłoby łatwe przeprowadzanie absurdalnych dowodów w rodzaju $4,0=4,1$.

Problem niedostrzegany przez wielu to tzw. *testowanie wielokrotne*. Jeżeli zakładamy prawdopodobieństwo błędnego odrzucenia hipotezy zerowej równe 0,05, to zgadzamy się, że przeciętnie raz na 20 decyzji popełniamy błąd. Ten poziom istotności dotyczy wszakże tylko „pojedynczego” testowania. Jeżeli stosujemy test niezależnie 20 razy, to prawdopodobieństwo, że przynajmniej raz popełnimy błąd nie wynosi co prawda (jak się niektórym wydaje) 1, ale nieco ponad $2/3$. Zagadnienia, w których mamy do czynienia z tego typu problemami, to porównywanie średnich parami (testy post-hoc w ANOVA), testowanie istotności elementów macierzy korelacji, budowa modeli regresji o dużej liczbie zmiennych objaśniających. W tych sytuacjach trzeba ocenić rozmiar „niebezpieczeństwa”



powodowanego przez testowanie wielokrotnie (prosto wynikające z liczby jednocześnie rozpatrywanych zmiennych lub grup), a następnie zastosować metody umożliwiające korektę poziomu istotności bądź wartości p .

Wartość p

Jest to element weryfikacji hipotez, ale wyróżniamy go w osobnym podpunkcie, gdyż nagromadzenie błędów jest tu wyjątkowo duże. Rozpoczynając już od samej nazwy. W języku angielskim jest to p value (niekiedy pisane przez duże P). Usiłowałem – z różnym skutkiem – wylansować kiedyś nazwę *prawdopodobieństwo testowe*, w analogii do statystyki testowej. Ta właśnie analogia jest prawdziwa, nie zaś analogia do poziomu istotności. *Wartość p* bywa często nazywana *zaobserwowanym poziomem istotności* lub *komputerowym poziomem istotności*. Nie jest oczywiście żadnym z nich. Czym więc jest? Podajmy tu trzy definicje:

1. Pole pod funkcją gęstości rozkładu prawdopodobieństwa statystyki testowej obliczone od empirycznej wartości tej statystyki w kierunku wskazanym przez hipotezę alternatywną. Pole to może być jednoczęściowe (przy jednostronnej hipotezie alternatywnej) lub dwuczęściowe (przy hipotezie dwustronnej).
2. Prawdopodobieństwo uzyskania wyniku bardziej przeczącego hipotezie zerowej niż ten wynik, który właśnie uzyskaliśmy.
3. Najostrzejszy poziom istotności, przy którym możemy odrzucić testowaną hipotezę na podstawie danych empirycznych, które posiadamy.

W klasycznym testowaniu decyzję o ewentualnym odrzuceniu hipotezy zerowej podejmujemy na podstawie wyniku porównania *empirycznej wartości statystyki testowej* z *wartością krytyczną* odczytaną z tablic rozkładu statystyki testowej. Identyczną decyzję możemy podjąć, porównując *wartość p* z *poziomem istotności α* . Ta reguła odrzucenia ($p \leq \alpha$) jest prawdziwa dla wszystkich testów statystycznych (nawet tych jeszcze nie wymyślonych) i nie wymaga wykorzystywania tablic statystycznych. Jej powszechne stosowanie jest możliwe dzięki programom komputerowym, które dla obliczenia pola nie muszą analitycznie wyznaczać całki z funkcji gęstości. Teraz badacz nie ma problemów obliczeniowych i powinien skoncentrować się na ważnych sprawach podstawowych: formułowanie hipotez, wybór testu, realność założeń, jakość danych statystycznych, interpretacja wyników.

Główna niewłaściwa interpretacja *wartości p* to uznawanie jej za prawdopodobieństwo prawdziwości hipotezy zerowej.

Wiele jest też niewłaściwości (lub przynajmniej braku elegancji) w prezentowaniu *wartości p* w publikacjach. Pole jest konkretną liczbą i zapis w postaci nierówności jest tu nie na miejscu. Przecież $p < 0,2756$ nic nie oznacza, bo w końcu nie wiadomo, czy to p jest mniejsze od $\alpha = 0,05$ czy nie. Nie należy obawiać się zapisu $p = 0,0000$. Przeciwnicy tego zapisu zapominają, że w statystyce w większości sytuacji *zero* nie oznacza *nic* (właściwym oznaczeniem jest tu kreska) tylko *bardzo mało*. Niemal powszechna w literaturze



światowej konwencji prezentowania wartości p z dokładnością do czwartego miejsca po przecinku uniemożliwia przedstawienia pola – powiedzmy – $0,00003$, inaczej niż $0,0000$. Maniera stosowania zapisu *NS* (*nonsignificant* – *nieistotny*) zamiast podawania wartości p jest niezrozumiałym ograniczeniem ważnej informacji. Przecież $p=0,8767$ (prawdopodobnie hipoteza zerowa jest prawdziwa) oznacza coś innego niż $p=0,1245$ (przy większej próbie jest szansa na udowodnienie istotności – warto szukać dalej), mimo że formalnie obydwie liczby podpadają pod kategorię *NS*.

Jak uniknąć błędów?

Przysłowie powiada, że należy uczyć się nie na błędach, tylko na uniwersytetach. Te uniwersytety możemy uogólnić na autorytety i dobre książki. To tam trzeba szukać porad i wskazówek.

Good i Hardin [1] wskazują na następujące źródła błędów popełnianych przy stosowaniu metod statystycznych:

- ◆ Używanie tego samego zbioru danych do formułowania i testowania hipotezy,
- ◆ Pobieranie próby z niewłaściwej populacji lub brak jej określenia przed badaniem,
- ◆ Próby, które są nielosowe lub niereprezentatywne,
- ◆ Pomiar złych zmiennych lub mierzenie nie tego, co chcieliśmy mierzyć,
- ◆ Użycie niewłaściwych metod statystycznych,
- ◆ Brak weryfikacji uzyskanych modeli,
- ◆ Pozwolenie na to, aby statystyczne procedury podejmowały decyzje za badacza.

Formułują oni częściową receptę na zastosowania statystyki wolne od błędów:

1. Sformułuj cele badań i sposób wykorzystania wyników, *zanim* rozpoczniesz eksperyment laboratoryjny, badanie kliniczne lub przygotowanie ankiety oraz *zanim* przeanalizujesz swój zbiór danych.
2. Określ populację, której mają dotyczyć wyniki Twoich badań.
3. Określ listę wszystkich możliwych źródeł wariacji. Kontroluj je lub mierz, aby ominąć ich związek z relacjami pomiędzy tymi zagadnieniami, które są głównym przedmiotem naszego zainteresowania.
4. Sformułuj hipotezy i wszystkie związane z nimi alternatywy. Określ możliwe wyniki eksperymentów, ich znaczenie i potencjalne wnioski. Zrób to, *zanim* zbierzesz jakiegokolwiek dane oraz *zanim* włączysz komputer.
5. Szczegółowo określ sposób pobierania próby.
6. Używaj właściwych estymatorów zgodnych, efektywnych, wystarczających, przedziałowych i odpornych.



7. Znaj założenia występujące w testach z których korzystasz. Używaj testów o ograniczonej liczbie założeń, ale mocnych (szczególnie względem alternatyw, które testujesz).
8. W sprawozdaniu z badań określ dokładnie badaną populację oraz sposób pobierania próby. Napisz, które elementy próby nie weszły do ostatecznego pliku danych i dlaczego.

Literatura

1. Good P.I., Hardin J.W., *Common Errors in Statistics (and How to Avoid Them)*, John Wiley & Sons, 2003.
2. Huffa D., *How To Lie With Statistics*, W.W. Norton & Company, 1954, 1982, 1993.