Wiadomości Statystyczne. The Polish Statistician, 2021, vol. 66, 3, 7–21 DOI: 10.5604/01.3001.0014.7984
Studia metodologiczne / Methodological studies
Otrzymano/received: 15.12.2020, zaakceptowano/accepted: 15.02.2021

Błędy nielosowe i ich znaczenie w testowaniu hipotez

Mirosław Szreder[a]

Streszczenie. We współczesnych badaniach reprezentacyjnych coraz częściej dają o sobie znać błędy o charakterze nielosowym, w tym w szczególności wynikające z braków odpowiedzi lub źle wykonanych pomiarów (niedokładnej obserwacji statystycznej). Do tej pory rzadko dyskutowano o skutkach tego typu błędów w procedurze weryfikacji hipotez statystycznych. Uwaga badaczy skupiała się niemal wyłącznie na błędzie losowania (błędzie losowym). Błąd ten maleje wraz ze wzrostem liczebności próby. To sprawia, że badacze, nierzadko mający do dyspozycji bardzo duże liczebnie próby, tracą z pola widzenia konsekwencje nie tylko błędu losowego, lecz także błędów nielosowych.

Celem artykułu jest wskazanie na znaczenie błędów nielosowych w podejmowaniu decyzji opartych na wykorzystaniu klasycznej procedury weryfikacji hipotez. Szczególną uwagę poświęcono sytuacjom, w których badacz dysponuje dużą liczebnie próbą. W pracy uzasadniono twierdzenie, że w dużych próbach testy statystyczne stają się bardziej wrażliwe na oddziaływanie błędów nielosowych. Błędy systematyczne, będące szczególnym przypadkiem błędów nielosowych, zwiększają prawdopodobieństwo błędnej decyzji o odrzuceniu prawdziwej hipotezy wraz ze wzrostem liczebności próby. Wzbogacenie weryfikacji hipotez o analizę opartą na estymacji przedziałowej może wspomóc badacza w poprawnym wnioskowaniu.

Słowa kluczowe: testowanie hipotez statystycznych, błąd losowania, błąd losowy, błędy nielosowe

JEL: C12, C13, C18, D80

Non-random errors and their importance in testing of hypotheses

Abstract. Increasing numbers of non-random errors are observed in contemporary sample surveying – in particular, those resulting from no response or faulty measurements (imprecise statistical observation). Until recently, the consequences of these kinds of errors have not been widely discussed in the context of the testing of hypoteses. Researchers focused almost entirely on sampling errors (random errors), whose magnitude decreases as the size of the random sample grows. In consequence, researchers who often use samples of very large sizes tend to overlook the influence random and non-random errors have on the results of their study.

The aim of this paper is to present how non-random errors can affect the decision-making process based on the classical hypothesis testing procedure. Particular attention is devoted to cases in which researchers manage samples of large sizes. The study proved the thesis that samples of large sizes cause statistical tests to be more sensitive to non-random errors. Systematic errors, as a special case of non-random errors, increase the probability of making the wrong decision to reject a true hypothesis as the sample size grows. Supplementing the testing of hypotheses with the analysis of confidence intervals may in this context provide substantive support for the researcher in drawing accurate inferences.

Keywords: testing of hypotheses, sampling error, random error, non-random errors

[a] Uniwersytet Gdański, Wydział Zarządzania, Katedra Statystyki / University of Gdansk, Faculty of Management, Department of Statistics. E-mail: miroslaw.szreder@ug.edu.pl.
ORCID: https://orcid.org/0000-0002-7597-0816.

© Główny Urząd Statystyczny

1. Wprowadzenie

Rzadko które metody i techniki statystyczne budzą w środowisku naukowym tyle kontrowersji, co statystyczne testy istotności (Dorofeev i Grant, 2006). Źródeł nieporozumień i polemik jest tu co najmniej kilka. Wiążą się one zarówno z niejednakowym rozumieniem w środowisku statystyków i poza nim kluczowego w testowaniu hipotez terminu *istotność*, jak i z brakiem konsensusu wśród części naukowców co do tego, w jakich warunkach mogą być stosowane testy i jak poprawnie należy interpretować ich rozstrzygnięcia (szerzej: Szreder, 2019).

Retrospektywne spojrzenie na ewolucję zastosowań testów statystycznych pozwala dostrzec, że wykorzystywane były one do realizacji dwóch odrębnych celów (Gigerenzer i Marewski, 2015). Sensem pierwszego – wcześniejszego z zastosowań, obecnego m.in. w fizyce i astronomii – było zaufanie do pewnej hipotezy, np. że rozkład błędów pomiarów wobec prawdziwego położenia określonej planety jest rozkładem normalnym, i odrzucenie obserwacji, które nazbyt odbiegały od tego rozkładu (najczęściej z winy obserwatora lub urządzeń pomiarowych). Celem tego rodzaju testowania było odrzucenie nie hipotezy, lecz niektórych obserwacji odstających (*outliers*). Drugie z zastosowań jest w swojej istocie niejako odwróceniem pierwszego. Uzyskane w próbie obserwacje uznaje się za godne zaufania, a celem testowania jest odrzucenie hipotezy, która okaże się zbyt odległa od tego, co w rzeczywistości zaobserwowano.

W klasycznym paradygmacie testowania hipotez statystycznych, zarówno w ujęciu Fishera, jak i Neymana-Pearsona, obserwacje uzyskane w próbie losowej nie podlegają ocenie ani weryfikacji, lecz służą stwierdzeniu, na ile nieprzystająca do tych obserwacji jest testowana hipoteza. Stopień nieprzystawania tej hipotezy (zerowej) do zaobserwowanej próby jest wyrażany w kategoriach probabilistycznych i stanowi o tym, czy hipoteza ta zostanie uznana za nieprawdziwą, a w konsekwencji – odrzucona. Najważniejszą rolę w rozstrzyganiu o nieprawdziwości hipotezy odgrywa więc próba statystyczna. A jeżeli tak, to w całej procedurze znaczenie mają zarówno błąd losowania, jak i błędy w dokonanych pomiarach, stanowiące jedną z kategorii błędów o charakterze nielosowym. O występowaniu i znaczeniu błędów nielosowych zdaje się zapominać część współczesnych badaczy i użytkowników metod statystycznych, którzy kładą nacisk niemal wyłącznie na pożądane wartości wskaźnika *p-value* i dążenie do uzyskania statystycznie istotnego rozstrzygnięcia. Trzeba podkreślić, że postawa twórców wnioskowania statystycznego była w tym zakresie zgoła odmienna. William S. Gosset – twórca rozkładu *t*-Studenta – uważał małe błędy pomiaru za ważniejsze od małych wartości *p-value* (Gigerenzer i in., 1989).

W naukach ekonomicznych i społecznych kategoria błędu pomiaru jest szczególnie pojemna, ponieważ obejmuje szereg uświadomionych i nieuświadomionych zniekształceń i pomyłek leżących zarówno po stronie realizatorów badania, w tym ankieterów, jak i po stronie respondentów. Nielosowy charakter, skutkujący często systematycznym obciążeniem, mają także inne błędy obecne w badaniach reprezentacyjnych, w tym błędy pokrycia i błędy braków odpowiedzi[1]. Warto więc rolę tych błędów oraz innych ułomności składających się na pojęcie błędów nielosowych rozważyć w kontekście współczesnego kryzysu replikowalności eksperymentów (w naukach ścisłych i społecznych), a szerzej – kryzysu statystycznej istotności[2].

Celem artykułu jest wskazanie na znaczenie błędów nielosowych w podejmowaniu decyzji opartych na wykorzystaniu klasycznej procedury weryfikacji hipotez. Szczególną uwagę poświęcono sytuacjom, w których badacz dysponuje dużą liczebnie próbą badawczą. Duża liczba obserwacji w próbie bywa współcześnie uznawana za usprawiedliwienie niespełnienia w praktyce empirycznej niektórych warunków wnioskowania statystycznego albo za remedium na ten stan rzeczy. Warto więc pokazać, jakie są skutki błędów losowych i nielosowych w testowaniu hipotez dla dużych prób. Do tej pory rzadko rozważano te konsekwencje w odniesieniu do błędów nielosowych.

Rozpocząć jednak trzeba od przypomnienia warunków, jakie powinny być spełnione, aby można było poprawnie stosować metody wnioskowania statystycznego, w tym weryfikacji hipotez. Najbardziej bowiem kontrowersyjną kwestią jest dziś coraz śmielsze dążenie do posługiwania się technikami wnioskowania statystycznego i interpretowanie jego wyników w kategoriach probabilistycznych, kiedy obserwacje w próbie nie mają charakteru losowego.

2. Znaczenie założenia o losowości próby

Łatwiejszy niż w przeszłości dostęp do danych statystycznych oraz do specjalistycznego oprogramowania komputerowego sprawia, że w wielu dziedzinach poznania naukowego wzrosło w ostatnich latach zainteresowanie metodami opisu statystycznego oraz wnioskowania. W szczególności testy statystyczne stały się jedną z najważniejszych grup narzędzi badawczych w różnych obszarach zastosowań nauk empirycznych. Wykorzystywane są do rozstrzygania na podstawie obserwacji w próbie o prawdziwości lub nieprawdziwości sformułowanych wcześniej hipotez odnoszących się do charakterystyk badanej populacji. Powszechność zastosowań technik wnioskowania statystycznego, w tym weryfikacji hipotez, świadczy z jednej strony

[1] O występowaniu tych kategorii błędów w badaniach statystyki publicznej pisze m.in. Paradysz (1989, 2009).

[2] Temat tego kryzysu obszernie – w ponad 40 artykułach – opisali statystycy z całego świata na łamach czasopisma „The American Statistician" w nr. 73 z marca 2019 r.

o potencjale i możliwościach tych technik, a z drugiej – przy masowym ich wykorzystaniu – staje się powodem uzasadnionych obaw o poprawność ich użycia. W literaturze światowej najszerzej w tym kontekście dyskutowana jest kwestia rozstrzygania za pomocą wskaźnika *p-value* o nieprawdziwości testowanej hipotezy, a także stosowania w tej procedurze stwierdzenia o *statystycznej istotności* efektu zaobserwowanego w próbie[3]. Jednak pierwotne, czyli wyjściowe, wobec tych kontrowersji jest inne zagadnienie, mianowicie sens wymogu losowości próby – jednego z najbardziej podstawowych założeń w teorii wnioskowania statystycznego.

Matematycznym modelem wnioskowania statystycznego jest – przypomnijmy – model probabilistyczny, który opiera się na założeniu o losowym mechanizmie generowania obserwacji w próbie. Z tym założeniem związane są najbardziej podstawowe i najważniejsze elementy estymacji i weryfikacji hipotez. Punktem wyjścia do zastosowań probabilistycznego modelu wnioskowania statystycznego jest uznanie, że nawet próba losowa nie stanowi doskonałego odzwierciedlenia struktury populacji, a w powtarzalnych losowaniach mogą występować znaczne różnice w charakterystykach otrzymywanych kolejno prób. Innymi słowy, jest to cechująca badacza świadomość istnienia błędu losowania (*sampling error*), nazywanego także błędem losowym (*random error*), który wyraża niemożność zapewnienia przez mechanizm generujący losowe obserwacje w próbie pełnej zgodności struktury próby ze strukturą populacji. Duże znaczenie tego błędu – który nierozerwalnie wiąże się z samym aktem losowania – we wnioskowaniu statystycznym wynika przynajmniej z trzech powodów.

Po pierwsze błąd losowania stanowi jedyny rodzaj błędu, który występuje w teoretycznym (matematycznym) modelu wnioskowania statystycznego. Brak w tym modelu innych elementów całkowitego błędu badania próbkowego (*total survey error*). W szczególności nieobecne są w nim błędy o charakterze nielosowym.

Po drugie błąd losowania – wraz z założeniami o rozkładach prawdopodobieństwa cech populacji, z której pochodzą losowe obserwacje – stanowi punkt wyjścia do określenia wszelkich miar precyzji i wiarygodności wnioskowania statystycznego. Dodatkowo od tego właśnie błędu (i tylko tego) rozpoczyna się procedura wyprowadzania formuł matematycznych i konstruowania narzędzi wnioskowania, które odgrywają kluczową rolę w uogólnianiu prawidłowości zaobserwowanych w próbie. Należą do nich przedziały ufności czy formuły na minimalną liczebność próby oraz najważniejsze statystyki związane z testowaniem hipotez, w tym *p-value*.

[3] Obu tym kwestiom poświęcają coraz więcej uwagi prestiżowe czasopisma naukowe, m.in.: „Science" (zob. Amrhein, Greenland i McShane, 2019) oraz „The American Statistician" (zob. Wasserstein i Lazar, 2016). Niektórzy autorzy piszą wręcz o kulcie istotności statystycznej, przy czym zaznaczają już na wstępie, że istotność statystyczna nie musi implikować istotności w podstawowym tego słowa znaczeniu, a niekiedy wręcz nie oznacza niczego, co byłoby warte uwagi (oryg. „Statistically significant relationships may, and often do, tell us nothing that matters"), o czym pisze Lempert (2009, s. 226). Zob. także: Gelman i Stern (2006), Ziliak i McCloskey (2008) oraz Szreder (2019).

Po trzecie oceny i decyzje mają we wnioskowaniu statystycznym właściwą, poprawną logicznie interpretację tylko wtedy, gdy odnoszą się do hipotetycznego ciągu powtarzalnych prób losowych i generowanego błędu losowania. Dotyczy to zarówno interpretacji przedziałów ufności, błędów pierwszego i drugiego rodzaju w testowaniu hipotez oraz wskaźnika *p-value*, jak i wszelkich właściwości estymatorów. Jeżeli badacz nie potrafi uzasadnić, że próba, którą dysponuje, jest losowa albo może być traktowana jak losowa, to wspomniane narzędzia wnioskowania statystycznego przestają zapewniać poprawną interpretację[4] (Hirschauer i in., 2020, s. 72)[5]. Podobnie dzieje się, gdy obserwacją są objęte wszystkie jednostki populacji, co w całości eliminuje błąd losowania w analizie statystycznej[6].

Wykorzystanie rachunku prawdopodobieństwa we wnioskowaniu statystycznym stwarza możliwość matematycznego opisania relacji pomiędzy próbą losową a populacją, którą ta próba reprezentuje. W naukach społecznych i eksperymentalnych za próbę losową stanowiącą podstawę wnioskowania uznaje się albo losowo uzyskany podzbiór jednostek ze zdefiniowanej wcześniej populacji (do której odnoszone będą późniejsze wnioski), albo zrandomizowany kontrolowany eksperyment, w którym jednostki poddane obserwacji trafiają losowo do dwóch grup: eksperymentalnej (poddanej oddziaływaniu jakiegoś bodźca) i kontrolnej[7].

Wydaje się, że jedną z głównych przyczyn niewystarczającej koncentracji badaczy na źródłach i konsekwencjach błędów nielosowych, w tym w szczególności błędów pomiaru, jest mylne przekonanie, że w warunkach łatwej dostępności dużych prób problem ten zanika albo wręcz nie istnieje. Duża liczebność próby zwodzi niektórych badaczy, upatrujących w niej gwaranta poprawnego i wiarygodnego wnioskowania niezależnie od jakości obserwacji w próbie. Najczęściej zapominają oni o tym, że najważniejsze kategorie błędów nielosowych nie są funkcjami liczebności próby. I nie tylko dlatego warto przeanalizować wzajemne relacje między liczebnością próby i błędami nielosowymi z jednej a poprawnością testowania hipotez z drugiej strony. Gigerenzer i Marewski (2015, s. 425) stwierdzają, że proste zwiększanie liczby

---

[4] Oryg. „they become essentially uninterpretable".

[5] Zob. także Vogt i in. (2014, s. 244). Autorzy stwierdzają, że w badaniach, w których nie korzysta się z randomizacji lub z próbkowania losowego, klasyczne podejście do wnioskowania statystycznego jest nieodpowiednie (oryg. „in research not employing random assignment or random sampling, the classical approach to inferential statistics is inappropriate").

[6] W tego typu sytuacjach (tj. w badaniach wyczerpujących) swoje znaczenie traci także termin *istotność statystyczna* w odniesieniu do parametrów populacji, np. różnic tych parametrów od zera. Źródłem tego terminu jest bowiem świadomość badacza, że w każdej zaobserwowanej próbie losowej tkwią konsekwencje błędu losowania. Statystycznie istotne różnice to takie, które nie dają się wyjaśnić oddziaływaniem wyłącznie błędu losowania, a ściślej – czyniłyby takie wyjaśnienie mało wiarygodnym (prawdopodobnym) w świetle dokonanych obserwacji. To uściślenie, dotyczące użycia prawdopodobieństwa w ocenie istotności statystycznej, jest szczególnie potrzebne w kontekście uzasadnionej, szerokiej krytyki dychotomizacji pojęcia *istotności statystycznej* – traktowania tego pojęcia zero-jedynkowo (binarnie).

[7] Gdy ścisła randomizacja jest niemożliwa w danym eksperymencie, zaleca się użycie takich technik, jak *propensity score* w celu zmniejszenia błędu wyboru (*selection bias*). Zob. np. Mercer i in. (2017).

obserwacji w próbie stało się środkiem zastępczym (substytutem) dążenia do minimalizowania błędów[8]. O ile jednak błąd losowania maleje wraz ze wzrostem liczby obserwacji w próbie, o tyle ważne jest, aby prawidłowości tej nie odnosić do całkowitego błędu wnioskowania, ponieważ jego pozostałe składniki, mające charakter nielosowy, nie podlegają tej prawidłowości.

Niedocenianiu błędów nielosowych sprzyja nie tylko złudne przekonanie o możliwościach sprawczych dużych prób, lecz także bardziej złożone zjawisko pomijania i lekceważenia wielu innych okoliczności w dążeniu do szybkiego osiągnięcia statystycznie istotnego wyniku. Ma to związek z błędnym wyobrażeniem odbiorców badań, a także niektórych naukowców, że wartościowe są tylko te wyniki, którym można przypisać etykietę statystycznie istotnych. W dążeniu do uzyskania statystycznie istotnych wyników pomija się zaś kwestię błędów nielosowych. Tymczasem te ostatnie mogą silniej od błędu losowania wpływać na rozstrzygnięcia o istotności różnic stanowiących podstawę do odrzucenia hipotezy zerowej.

3. Testowanie hipotez na podstawie liczebnie dużych prób

Problem błędów nielosowych w testowaniu hipotez nie ujawniłby się prawdopodobnie z taką ostrością, gdyby nie to, że coraz częściej badacze mają dostęp do liczebnie dużych prób, znacznie przekraczających wyobrażenia twórców teorii wnioskowania statystycznego z początków XX w. Wykorzystanie wielotysięcznej próby do weryfikacji hipotez statystycznych rodzi różne konsekwencje, z których tylko część była do tej pory wyraźnie komunikowana. Najczęściej skupiano się na poprawie mocy testu wraz ze wzrostem liczebności próby. Innymi słowy, stwierdzano słusznie, że rosnąca liczebność próby zwiększa zdolność testu statystycznego do prawidłowego rozróżnienia pomiędzy hipotezą prawdziwą a fałszywą. Raz jeszcze warto zaznaczyć, że moc testu (dopełnienie do 1 prawdopodobieństwa błędu drugiego rodzaju) nie uwzględnia żadnych innych – poza błędem losowania – błędów, którymi mogą być obciążone wyniki w próbie.

Pozostańmy na razie przy powyższym założeniu, abstrahując od oddziaływania błędów nielosowych. Już tutaj można dostrzec pierwsze wyzwania, przed jakimi staje statystyk dysponujący dużą próbą losową. Warto zaznaczyć, że po raz pierwszy wskazywano na nie już ponad pół wieku temu. W 1966 r. na łamach „Psychological Bulletin" David Bakan trafnie uzasadniał, że każda hipoteza zerowa może zostać odrzucona, jeżeli tylko wylosuje się odpowiednio dużą próbę. Zilustrował to m.in. doświadczeniem, w którym sam uczestniczył, polegającym na przetestowaniu zwykłych baterii przez 60 tys. Amerykanów. Za pomocą podziału tej próby na dwie gru-

[8] Oryg. „Simply increasing the number N of subjects became a surrogate for minimizing errors".

py osób dokonywanego według zupełnie dowolnych i nic nieznaczących kryteriów, takich jak mieszkanie po wschodniej lub zachodniej stronie rzeki Missisipi, na północy lub południu kraju, uzyskiwał za każdym razem statystycznie istotne różnice w średnich. I nie były to w tamtym czasie jedyne spostrzeżenia prowadzące do wniosku, że „jeżeli hipoteza zerowa nie zostaje odrzucona, to zwykle z tego powodu, że liczebność próby (n) była za mała”[9] (Nunnally, 1960, s. 643). W znanej monografii z końca lat 70. XX w. Edward E. Leamer pisał bardziej dosadnie, że w dużych próbach nawet bardzo mała i w danym zagadnieniu nieznacząca wielkość efektu będzie prowadzić do stwierdzenia jego statystycznej istotności (czyli odrzucenia hipotezy zerowej), „ponieważ duża próba jest przypuszczalnie bardziej miarodajna niż mała, a hipotezę zerową odrzucimy dla odpowiednio dużej próby, moglibyśmy równie dobrze tę hipotezę odrzucić bez pobierania próby w ogóle”[10] (1978, s. 89).

Opisane wyżej prawidłowości można uzasadnić także analitycznie, odwołując się do najbardziej popularnego obecnie wskaźnika w procedurze weryfikacji hipotez – *p-value*. Lin i in. (2013) prezentują matematyczny dowód na to, że wartości *p-value*, oparte na zgodnych estymatorach, charakteryzują się następującą własnością asymptotyczną w odniesieniu do hipotezy zerowej H_0: $\beta = 0$:

$$\lim_{n\to\infty} p\text{-}value(n) = \lim_{n\to\infty} P_n\left(|\hat{\beta} - \beta| < \varepsilon\right) = \begin{cases} 0 \text{ dla } \beta \neq 0 \\ 1 \text{ dla } \beta = 0 \end{cases}.$$

Oznacza to, że przy wzrastającej do nieskończoności liczebności próby (n) cała masa prawdopodobieństwa w rozkładzie próbkowym zgodnego estymatora $\hat{\beta}$ coraz ściślej skupia się wokół parametru populacji β. Jeżeli w populacji parametr β jest dokładnie równy hipotetycznej wartości 0, wówczas dla dużych liczebności prób rozkład estymatora $\hat{\beta}$ jest tak silnie skoncentrowany wokół 0, że bliskie pewności jest otrzymanie próby, w której $\hat{\beta} = 0$. To z kolei oznacza – przy wzięciu pod uwagę interpretacji *p-value* – że w tej sytuacji wraz ze wzrostem liczebności próby wartości *p-value* rosną do 1. W pozostałych przypadkach, kiedy parametr populacji odbiega od 0 (od wartości hipotetycznej), choćby bardzo nieznacznie (np. na dalekim miejscu po przecinku), to konsekwentnie trzeba zauważyć, że właśnie w tym punkcie – bliskim, ale nie równym 0 – skupia się niemal całe prawdopodobieństwo rozkładu próbkowego estymatora $\hat{\beta}$. To oznacza, że dla $n \to \infty$ zbiega do 1 prawdopodobieństwo otrzymania próby, dla której $\hat{\beta} = \beta \neq 0$, a do 0 – prawdopodobieństwo, które wyraża wskaźnik *p-value* – otrzymania zaobserwowanej próby przy założeniu, że

[9] Oryg. „If the null hypothesis is not rejected, it usually is because the N is too small”. Tłumaczenie tego i wszystkich pozostałych cytatów w artykule pochodzi od autora.

[10] Oryg. „A large sample is presumably more informative than a small one, and since it is apparently the case that we will reject the null hypothesis in a sufficiently large sample, we might as well begin by rejecting the hypothesis and not sample at all”.

prawdziwa jest hipoteza zerowa $\beta = 0$. Innymi słowy, z wyjątkiem rzadkich w praktyce badawczej sytuacji, kiedy parametr β jest dokładnie – co do dalekich miejsc po przecinku – równy wartości ujętej w hipotezie zerowej (wtedy *p-value* zbiega do 1), wartości wskaźnika *p-value* są zbieżne do 0 przy założeniu rosnącej do nieskończoności liczebności próby.

W realnych populacjach poddawanych analizom i wnioskowaniu testowany parametr bardzo rzadko bywa dokładnie równy tej wartości, jaką zapisano w hipotezie zerowej. Dlatego liczebnie duże próby będą prawie zawsze dawały wartości *p-value* bliskie 0, skutkujące odrzuceniem hipotezy zerowej. Z wyraźną łatwością odrzucane są np. hipotezy o tym, że współczynniki korelacji czy regresji są równe 0, jeżeli tylko badacz posiada odpowiednio dużą próbę losową. W populacji bowiem niesłychanie rzadko wartości tych współczynników są dokładnie równe 0 (jak głosi hipoteza zerowa). Badacz mający do dyspozycji dużą próbę bez trudu odrzuci hipotezę głoszącą, że testowany parametr jest statystycznie nieistotny (równy 0), mimo że w rzeczywistości jego wartość będzie mogła być bardzo bliska 0. Prowadzi to oczywiście do przykrych błędów poznawczych – rozbieżności między istotnością naukową a istotnością statystyczną. Tę kłopotliwą konkluzję dla zastosowań testów w dużych próbach wyraził dobitnie m.in. Cohen (1990, s. 1308):

> Hipoteza zerowa traktowana dosłownie (a tylko tak można ją rozumieć w formalnej procedurze weryfikacji hipotez) jest zawsze fałszywa w realnej rzeczywistości... Jeżeli zaś jest fałszywa, choćby w bardzo nieznacznym stopniu, to z odpowiednio dużej próby uzyska się wynik istotny statystycznie, który doprowadzi do jej odrzucenia[11].

Ta decyzja z kolei oznacza, że dla dużych prób wiele nieznaczących różnic i określonych w hipotezie zerowej relacji między zmiennymi zostanie uznanych za istotne statystycznie. Według Lemperta (2009, s. 230) „przy odpowiednio dużej wielkości próby niemal wszystkie zależności w próbie będą istotne statystycznie, ponieważ dla rosnącej liczebności próby znikają efekty losowe i nawet słabe sygnały przebiją się przez szum realnego świata”[12]. Przez *sygnał* należy rozumieć wartość statystyki testowej w próbie, wyrażającej niedopasowanie uzyskanej próby do brzmienia hipotezy zerowej. Natomiast desygnatem *szumu* – albo zakłóceń – jest wartość miary rozproszenia statystyki testowej, np. odchylenie standardowe. W dużych próbach, jak wskazano wcześniej, należy oczekiwać niewielkich zakłóceń (małych wartości odchy-

[11] Oryg. „The null hypothesis, taken literally (and that's the only way you can take it in formal hypothesis testing), is always false in the real world... If it is false, even to a tiny degree, it must be the case that a large enough sample will produce a significant result and lead to its rejection”.

[12] Oryg. „With a large enough N, virtually all associations in a sample will be statistically significant, for as size increases, random effects are more likely to cancel out, and even weak signals will emerge through the real world's noise”.

lenia standardowego statystyki testowej), stąd nawet niewielka niezgodność próby z wynikami, jakie powinny się były pojawić, gdyby prawdziwa była hipoteza zerowa, doprowadzi do odrzucenia tej hipotezy.

W tym kontekście rzadko się do tej pory zwracało uwagę na błędy o charakterze nielosowym, których oddziaływanie dodatkowo zwiększa szanse odrzucenia hipotezy zerowej w dużych próbach. Dotyczy to w szczególności sytuacji, gdy testowana hipoteza (np. o statystycznie nieistotnej różnicy parametru populacji od 0) nie powinna zostać uznana za nieprawdziwą i odrzucona. Znaczenie tej kategorii błędów (nielosowych) wyraźnie wzrosło we współczesnych badaniach statystycznych[13], co wciąż pozostaje niedocenione w światowej literaturze statystycznej i w praktyce badań empirycznych. Efektem oddziaływania błędów nielosowych jest najczęściej obciążenie uzyskanych ocen błędem systematycznym. Co prawda wielkość tego błędu nie jest funkcją wielkości próby, ale jego wpływ na poprawność wnioskowania (słuszność odrzucenia testowanej hipotezy) wykazuje taką zależność.

Gdy systematyczne obciążenie jest nieduże, a liczebność próby niewielka, to ryzyko błędnej decyzji o odrzuceniu hipotezy zerowej, gdy jest ona prawdziwa, nie jest duże. Przed tego typu błędnymi decyzjami chroni duże rozproszenie statystyki testowej, wyrażające znaczną niepewność badacza. Obszar nieodrzucenia hipotezy zerowej jest w tym przypadku spory. Ryzyko to oznacza, w kontekście wskaźnika *p-value*, że gruby ogon w rozkładzie statystyki testowej czyni dość prawdopodobnym przyjęcie przez nią wartości odległych od tej, która wynikałaby z hipotezy zerowej (co odpowiada stosunkowo dużym wartościom *p-value*). W rezultacie hipoteza zerowa nie zostanie odrzucona.

Systematyczny błąd staje się groźniejszy wtedy, gdy towarzyszy mu duża liczebność próby. Statystyczna prawidłowość, silnie potwierdzona przez dużą wielkość próby – a w rezultacie przez małe rozproszenie statystyki testowej – jest trudna do zakwestionowania. Rzadko się jednak podkreśla, że każda taka prawidłowość jest w dużym stopniu warunkowa, ponieważ zachodzi jedynie wtedy, gdy obserwacje w próbie nie są obciążone błędami nielosowymi, w szczególności systematycznymi. Istnienie błędu systematycznego w dokonanych pomiarach w próbie – przy małym lub bardzo małym rozproszeniu statystyki testowej – sprawia, że obszar odrzucenia hipotezy zerowej znacznie się rozszerza. Innymi słowy, jakakolwiek wartość statystyki różna od tej, która wynika z brzmienia hipotezy zerowej, staje się bardzo mało prawdopodobna. Uzyskanie takiej lub bardziej skrajnej wartości (*p-value*) jest bliskie 0. Błąd systematyczny będzie w tego typu sytuacjach zwiększał ryzyko niesłusznego odrzucenia hipotezy zerowej. Im większa jest liczebność próby, tym bardziej wrażliwy na błędy systematyczne staje się test statystyczny. Co prawda niewielki błąd tego

[13] O zmianach w strukturze całkowitego błędu badania próbkowego pisze m.in. Szreder (2015).

rodzaju oznacza niewielkie przesunięcie rozkładu statystyki testowej względem osi rzędnych, ale przy bardzo małej dyspersji rozkładu tej statystyki ogromna część masy prawdopodobieństwa przesuwa się w kierunku nieprawdziwej wartości. W rozkładach bardziej rozproszonych, odpowiadających mniejszym liczebnościom próby, jest to mniejsza frakcja masy prawdopodobieństwa pod krzywą rozkładu. Zależności te ilustruje poniższy przykład.

Załóżmy, że weryfikacji na poziomie istotności 0,05 podlega hipoteza stwierdzająca, że średnia μ w populacji o rozkładzie normalnym, w którym odchylenie standardowe wynosi 4, równa jest 5:

$$H_0: \mu = 5$$
$$H_1: \mu \neq 5$$

Oznacza to, że jeżeli rzeczywiście średnia wartość w tej populacji wynosi 5, to hipotezę zerową odrzucimy w tych próbach losowych, dla których wartość statystyki testowej

$$Z = \frac{\bar{X} - \mu}{\sigma} \cdot \sqrt{n}$$

przyjmie wartości mniejsze od –1,96 lub większe od +1,96, gdzie $\bar{X}$ oznacza średnią arytmetyczną z próby, σ – odchylenie standardowe w populacji (równe 4), a n – liczebność próby. W tych przypadkach zmaterializuje się błąd pierwszego rodzaju (odrzucenie prawdziwej hipotezy), a prawdopodobieństwo jego popełnienia wyniesie 0,05. Innymi słowy, jeżeli wnioskowanie wolne jest od innych błędów – w szczególności nielosowych – to przeciętnie 5 na 100 wylosowanych prób będzie prowadzić do błędnej decyzji o odrzuceniu prawdziwej hipotezy zerowej.

Z kolei aby zilustrować skutki błędów nielosowych – a zwłaszcza konsekwencje błędu systematycznego – można obliczyć, jak rośnie prawdopodobieństwo podjęcia błędnej decyzji o odrzuceniu prawdziwej hipotezy zerowej, gdy obserwacje w próbie obciążone są takim właśnie błędem. Prawdopodobieństwo to jest wyraźnie większe od 0,05 i rośnie wraz ze wzrostem wielkości próby. Obliczenia obrazujące, jak się ono zmienia, wykonano dla dwóch wielkości błędu systematycznego, wynoszącego:

a) +0,5 (wszystkie dokonane pomiary w próbie są zawyżone o 0,5 jednostki pomiaru);
b) +1,0 (wszystkie pomiary są zawyżone o 1 jednostkę pomiaru).

Dla ustalonych liczebności prób n = 16, 36, 80, 100, 200, 500 zostały obliczone prawdopodobieństwa tego, że wartości zmiennej losowej $\bar{X} \sim N\left(5{,}5; {}^{4}/_{\sqrt{n}}\right)$ w przy-

padku a) oraz zmiennej losowej $\bar{X} \sim N\left(6;\ {}^{4}/_{\sqrt{n}}\right)$ w przypadku b) przyjmą wartości z obszaru krytycznego – przedziału odrzucenia hipotezy zerowej (kolumna 2 w poniższej tablicy).

Tablica. Prawdopodobieństwa błędnej decyzji o odrzuceniu hipotezy zerowej w zależności od wielkości błędu systematycznego i liczebności próby

Liczebność próby (*n*)	Obszar krytyczny – przedział średniej z próby prowadzący do odrzucenia hipotezy zerowej	Prawdopodobieństwo błędnej decyzji o odrzuceniu hipotezy zerowej	
		błąd systematyczny = 0,5	błąd systematyczny = 1,0
16	$(-\infty;\ 3{,}04) \cup (6{,}96;\ +\infty)$	0,079	0,170
36	$(-\infty;\ 3{,}69) \cup (6{,}31;\ +\infty)$	0,117	0,323
80	$(-\infty;\ 4{,}12) \cup (5{,}88;\ +\infty)$	0,201	0,609
100	$(-\infty;\ 4{,}22) \cup (5{,}78;\ +\infty)$	0,240	0,705
200	$(-\infty;\ 4{,}45) \cup (5{,}55;\ +\infty)$	0,424	0,942
500	$(-\infty;\ 4{,}65) \cup (5{,}35;\ +\infty)$	0,798	≈1

Źródło: obliczenia własne.

W tym przykładzie jest widoczna wyraźna tendencja do wzrostu prawdopodobieństwa podjęcia błędnej decyzji o odrzuceniu testowanej hipotezy wraz ze wzrostem liczebności próby. Im mniejszy błąd systematyczny, tym ryzyko błędnej decyzji jest mniejsze.

Wspomniana wcześniej warunkowość prawidłowości statystycznych zaobserwowanych w próbie, odnosząca się do nieobciążenia obserwacji błędami systematycznymi, powinna być podkreślana każdorazowo przy interpretacji *p-value*. Wskaźnik ten wyraża prawdopodobieństwo uzyskania takiej próby, jaką zaobserwowano, albo bardziej ekstremalnej (bardziej nieprzystającej do hipotezy zerowej), pod warunkiem że nie tylko hipoteza zerowa jest prawdziwa, lecz także spełnione są wszystkie inne założenia w modelu wnioskowania[14]. W szczególności obejmują one losowe generowanie obserwacji w próbie oraz niewystępowanie błędów systematycznych w dokonanych pomiarach. Często jednak o tej drugiej części założeń zapomina się nie tylko przy interpretacji *p-value*, lecz także w całej procedurze wnioskowania, w której przez długi czas nie przywiązywano wagi do błędów nielosowych.

Kwestia wrażliwości testów statystycznych na błędy nielosowe w dużych próbach powoduje, że ta wrażliwość staje się jeszcze jednym argumentem na rzecz wzbogacenia testowania hipotez o inne elementy wnioskowania statystycznego. Przede wszystkim warto w tym kontekście wspomnieć o przedziałach ufności, które najczę-

[14] Zwraca się na to uwagę m.in. w oświadczeniu Amerykańskiego Towarzystwa Statystycznego na temat istotności statystycznej oraz *p-value*, zob. Wasserstein i Lazar (2016, s. 131 i 132), a także Szreder (2019).

ściej dostarczają badaczowi więcej informacji niż tylko wskazanie, czy hipotezę dotyczącą danego parametru populacji należy odrzucić, czy uznać, że brak jest podstaw do jej odrzucenia. Jest to tym ważniejsze, że decyzja o braku podstaw do uznania hipotezy za fałszywą bywa niekiedy interpretowana implicite jako uznanie jej za prawdziwą. W celu zilustrowania tego, jakich informacji dostarcza badaczowi zastosowanie procedury weryfikacji hipotez, a jakich estymacja przedziałowa, rozważymy następujący przykład. Przyjmijmy, że testowana jest hipoteza o równości średnich w dwóch populacjach:

$$H_0: \mu_1 - \mu_2 = 0$$
$$H_1: \mu_1 - \mu_2 \neq 0$$

gdzie: μ_1 oznacza średnią arytmetyczną w pierwszej, a μ_2 – analogiczną średnią w drugiej populacji.

Załóżmy, że na podstawie wyników prób wylosowanych z tych populacji oszacowano 95-procentowy przedział ufności dla różnicy $\mu_1 - \mu_2$, który przybrał postać [1, 5]. Oznacza to, że badacz odrzuci sformułowaną wyżej hipotezę zerową na poziomie istotności 0,05, gdyż wartość 0 (ujęta w hipotezie zerowej) nie mieści się w przedziale liczbowym [1, 5]. Punktem centralnym tego przedziału, będącym próbkową oceną różnicy między średnimi w obu populacjach, jest wartość 3. Przyjmijmy następnie, że inna próba o tej samej liczebności dała tę samą wartość 3 jako ocenę różnicy między średnimi w obu populacjach, lecz rozproszenie wyników w próbie jest teraz większe, co powoduje, że przedział ufności ma postać [–1, 7]. Dla takiej próby zastosowanie procedury weryfikacji hipotez prowadzi do innej decyzji – mianowicie braku podstaw do odrzucenia hipotezy, że średnie w obu populacjach są identyczne. Jest tak dlatego, że wartość 0 mieści się w przedziale [–1, 7]. Czyli w tej drugiej sytuacji, mimo że różnica między średnimi próbkowymi wynosi 3 i jest taka, jak w pierwszym przypadku, badacz uzna, że nie ma podstaw do odrzucenia hipotezy zerowej. Dla niektórych odbiorców takiego komunikatu staje się to podstawą do uznania, że w dalszym procesie badawczym można założyć, że średnie te są równe. Tymczasem przedziały ufności znacznie lepiej obrazują to, co w przypadku weryfikacji hipotez kryje się za dychotomicznym wyborem – odrzucenia sprawdzanej hipotezy albo uznania, że brak jest podstaw do jej odrzucenia. W przypadku drugiej z rozważanych prób suchy komunikat z zastosowania testowania hipotez pozostawi wielu odbiorców w przekonaniu o dopuszczalności przyjęcia w przybliżeniu, że średnie w obu populacjach są identyczne. Natomiast przedział ufności sugeruje, że warto dokładnie to zbadać, ponieważ co prawda zerowa różnica między średnimi w populacjach nie jest – w świetle wyników próby – wykluczona, ale istnieje większe

prawdopodobieństwo, że różnica ta jest różna od 0 i dodatnia. Z tego i podobnych powodów coraz więcej czasopism naukowych oczekuje wzbogacenia testowania hipotez o przedstawienie przedziałów ufności, które pokazują pełny zakres różnic między tym, co zaobserwowano w próbie, a tym, co zapisano w hipotezie zerowej[15]. Ma to znaczenie także w sytuacjach, gdy dopuszcza się możliwość obciążenia danych w próbie błędami systematycznymi. Analiza przedziału ufności ogranicza ryzyko podjęcia błędnej decyzji opartej na procedurze testowania hipotez.

4. Podsumowanie

W empirycznych zastosowaniach wnioskowania statystycznego, w szczególności w badaniach ekonomicznych i społecznych, najbardziej uciążliwe są nie błędy losowe – których wielkość potrafimy wyrazić i zinterpretować – ale błędy nielosowe. Konsekwencje takich błędów, jak odmowy respondentów, braki w operacie losowania (błędy pokrycia), błędy treści czy powstałe na dalszych etapach badania błędy przetwarzania danych, stanowią współcześnie największe wyzwania dla statystyków. Wielu badaczy, świadomych tych błędów, usiłuje przekonać odbiorców wyników, że odpowiednio duża próba niweluje lub unieważnia konsekwencje ich oddziaływania. W artykule pokazano, że tak nie jest.

Testy statystyczne stosowane w dużych próbach losowych stają się bardziej niż w małych próbach wrażliwe zarówno na błędy losowe, jak i nielosowe. Badacz mający do dyspozycji dużą liczbę obserwacji w próbie bez trudu odrzuci hipotezę o nieistotności parametru populacji nawet wówczas, gdy w rzeczywistości jego wartość będzie bardzo bliska 0. Niewielka dyspersja statystyki testowej dla dużych prób sprawia bowiem, że dowolnie mała niezgodność próby z wynikami, jakie powinny się były pojawić, gdyby prawdziwa była hipoteza zerowa, doprowadzi do odrzucenia tej hipotezy. Jest to tylko kwestia odpowiedniej wielkości próby.

Wrażliwość ta odnosi się także do błędów o charakterze nielosowym, szczególnie błędów systematycznych. W dużych próbach każde obciążenie systematyczne w dokonanych obserwacjach skutkuje znacznym wzrostem ryzyka odrzucenia testowanej hipotezy w sytuacji, gdy w rzeczywistości jest ona prawdziwa. Ponadto im większa liczebność próby obarczonej błędem systematycznym, tym większe prawdopodobieństwo odrzucenia prawdziwej hipotezy zerowej. W artykule pokazano, że jednym ze sposobów zmniejszenia ryzyka błędnych decyzji w testowaniu hipotez może być analiza przedziału ufności przeprowadzona równolegle lub zamiast weryfikacji hipotez.

[15] Amrhein, Trafimow i Greenland (2019, s. 266) postulują: „Interpretuj i zwracaj uwagę raczej na otrzymane oceny niż na testy, uważnie analizując wartości pomiędzy dolną i górną granicą przedziału ufności".

Bibliografia

Amrhein, V., Greenland, S., McShane, B. (2019). Retire statistical significance. *Nature*, (567), 305–307. https://media.nature.com/original/magazine-assets/d41586-019-00857-9/d41586-019-00857-9.pdf.

Amrhein, V., Trafimow, D., Greenland, S. (2019). Inferential Statistics as Descriptive Statistics: There Is No Replication Crisis if We Don't Expect Replication. *The American Statistician*, *73*(sup1), 262–270. https://doi.org/10.1080/00031305.2018.1543137.

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, *66*(6), 423–437. https://doi.org/10.1037/h0020412.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, *45*(12), 1304–1312. https://doi.org/10.1037/0003-066X.45.12.1304.

Dorofeev, S., Grant, P. (2006). *Statistics for Real-Life Sample Surveys. Non-Simple-Random Samples and Weighted Data*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511543265.

Gelman, A., Stern, H. (2006). The Difference Between "Significant" and "Not Significant" is not Itself Statistically Significant. *The American Statistician*, *60*(4), 328–331. https://doi.org/10.1198/000313006X152649.

Gigerenzer, G., Marewski, J. N. (2015). Surrogate Science: The Idol of a Universal Method for Scientific Inference. *Journal of Management*, *41*(2), 421–440. https://doi.org/10.1177/0149206314547522.

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., Krüger, L. (1989). *The Empire of Chance: How probability changed science and everyday life*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511720482.

Hirschauer, N., Grüner, S., Mußhoff, O., Becker, C., Jantsch, A. (2020). Can *p*-values be meaningfully interpreted without random sampling?. *Statistics Surveys*, *14*, 71–91. https://doi.org/10.1214/20-SS129.

Leamer, E. E. (1978). *Specification Searches: Ad Hoc Inference with Nonexperimental Data*. New York: John Wiley & Sons. https://www.anderson.ucla.edu/faculty_pages/edward.leamer/books/specification_searches.htm.

Lempert, R. O. (2009). The Significance of Statistical Significance: Two Authors Restate An Incontrovertible Caution. Why a Book?. *Law & Social Inquiry*, *34*(1), 225–249. https://doi.org/10.1111/j.1747-4469.2009.01144.x.

Lin, M., Lucas, H. C., Jr., Shmueli, G. (2013). Too Big to Fail–Large Samples and the p-Value Problem. *Information Systems Research*, *24*(4), 906–917. https://doi.org/10.1287/isre.2013.0480.

Mercer, A. W., Kreuter, F., Keeter, S., Stuart, E. A. (2017). Theory and practice in nonprobability surveys. Parallels between causal inference and survey inference. *Public Opinion Quarterly*, *81*(S1), 250–279. https://doi.org/10.1093/poq/nfw060.

Nunnally, J. (1960). The place of statistics in psychology. *Educational and Psychological Measurement*, *20*(4), 641–650. https://doi.org/10.1177/001316446002000401.

Paradysz, J. (1989). O błędach nielosowych w badaniu dzietności kobiet w ramach Narodowego Spisu Powszechnego 1970. W: Główny Urząd Statystyczny, *Problemy badań statystycznych metodą reprezentacyjną* (s. 154–159). Warszawa.

Paradysz, J. (2009). Błędy pokrycia w Narodowych Spisach Powszechnych. Statystyka w praktyce społeczno-gospodarczej. *Prace Naukowe / Akademia Ekonomiczna w Katowicach*, 65–76.

Szreder, M. (2015). Zmiany w strukturze całkowitego błędu badania próbkowego. *Wiadomości Statystyczne*, *60*(1), 4–12.

Szreder, M. (2019). Istotność statystyczna w czasach big data. *Wiadomości Statystyczne. The Polish Statistician*, *64*(11), 42–57. https://ws.stat.gov.pl/Article/2019/11/042-057.

Vogt, W. P., Vogt, E. R., Gardner, D. C., Haeffele, L. M. (2014). *Selecting the Right Analyses for Your Data: Quantitative, Qualitative, and Mixed Methods*. New York: The Guilford Press.

Wasserstein, R. L., Lazar, N. A. (2016). The ASA's Statement on *p*-Values: Context, Process, and Purpose. *The American Statistician*, *70*(2), 129–133. https://doi.org/10.1080/00031305.2016.1154108.

Ziliak, S. T., McCloskey, D. N. (2008). *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*. Ann Arbor: University of Michigan. https://doi.org/10.3998/mpub.186351.