Wiadomości Statystyczne. The Polish Statistician, 2022, vol. 67, 8, 1–16 DOI: 10.5604/01.3001.0015.9704
Studia metodologiczne / Methodological studies
Zgłoszony/submitted: 4.05.2022, poprawiony/revised: 8.08.2022, zaakceptowany/accepted: 12.08.2022

Szanse i iluzje dotyczące korzystania z dużych prób we wnioskowaniu statystycznym

Mirosław Szreder[a]

Streszczenie. Teoria wnioskowania statystycznego jasno określa korzyści związane z dużą liczebnością próby badawczej. Wraz ze wzrostem wielkości próby maleje ilość błędów ocen szacowanych parametrów populacji (zwiększa się precyzja estymacji), a także rosną wartości mocy testów wykorzystywanych do weryfikacji hipotez statystycznych. Współczesne możliwości łatwego dotarcia do dużych prób badawczych (np. paneli internetowych), a także korzystania z coraz bardziej zaawansowanego i przyjaznego dla użytkownika oprogramowania statystycznego sprzyjają niedostrzeganiu zagrożeń dla wnioskowania statystycznego, jakie wiążą się z dużymi liczebnie próbami. Część badaczy ulega iluzji, że duża próba jest w stanie zniwelować i rozproszyć nie tylko błąd losowy, charakterystyczny dla każdej techniki losowania próby, lecz także błędy nielosowe. Znaczenie dużej liczebności próby jest ponadto jednym z ważnych aspektów toczącej się od kilkunastu lat dyskusji na temat istotności statystycznej (*p-value*) oraz problemów z jej rozstrzyganiem i interpretowaniem.

Celem opracowania jest wskazanie i omówienie konsekwencji dostrzegania w dużych próbach statystycznych jedynie szans, a pomijanie wyzwań i zagrożeń wynikających z ich stosowania. W artykule pokazano, że duża liczebność próby, której doboru dokonano za pomocą techniki nieprobabilistycznej, nie może stanowić alternatywy dla wyboru losowego. W szczególności dotyczy to internetowych paneli wolontariuszy deklarujących chęć udziału w badaniu. Wskazano ponadto na znaczenie komponentu nielosowego w błędzie próbkowania, który nie jest malejącą funkcją liczebności próby. W odniesieniu zaś do współczesnych problemów weryfikacji hipotez nakreślono i zilustrowano przykładem naukowy i etyczny wymiar podążania za istotnością statystyczną z wykorzystaniem dużych liczebnie prób lub wielokrotnego próbkowania.

Słowa kluczowe: wnioskowanie statystyczne, błąd próbkowania, błąd losowy, liczebność próby, istotność statystyczna, *p-value*

JEL: C12, C13, C18, D80

Opportunities and illusions of using large samples in statistical inference

Abstract. The theory of statistical inference clearly describes the benefits of large samples. The larger the sample size, the fewer standard errors of the estimated population parameters (the precision of the estimation improves) and the values of the power of statistical tests in hypothesis testing increase. Today's easy access not only to large samples (e.g. web panels) but also to more advanced and user-friendly statistical software may obscure the potential threats faced by statistical inference based on large samples. Some researchers seem to be under the illusion

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

that large samples can reduce both random errors, typical for any sampling technique, as well as non-random errors. Additionally, the role of a large sample size is an important aspect of the much discussed in the recent years issue of statistical significance (*p*-value) and the problems related to its determination and interpretation.

The aim of the paper is to present and discuss the consequences of focusing solely on the advantages of large samples and ignoring any threats and challenges they pose to statistical inference. The study shows that a large-size sample collected using one of the non-random sampling techniques cannot be an alternative to random sampling. This particularly applies to online panels of volunteers willing to participate in a survey. The paper also shows that the sampling error may contain a non-random component which should not be regarded as a function of the sample size. As for the contemporary challenges related to testing hypotheses, the study discusses and exemplifies the scientific and ethical aspects of searching for statistical significance using large samples or multiple sampling.

Keywords: statistical inference, sampling error, random error, sample size, statistical significance, *p*-value

1. Wprowadzenie

Przez dziesiątki lat zagadnienie dużych liczebnie prób kojarzyło się przede wszystkim z korzyściami dla wiarygodności i precyzji wnioskowania statystycznego. Korzyści te są bezdyskusyjne i dotyczą wielu ważnych elementów matematycznego modelu wnioskowania statystycznego. Są one widoczne zarówno w samych podstawach modelu wnioskowania, jak i w procedurach estymacji i weryfikacji hipotez, przede wszystkim w stosowanej w statystyce częstościowej interpretacji prawdopodobieństwa, w której duża liczba obserwacji (doświadczeń) zwiększa dokładność pomiaru prawdopodobieństwa[1]. Duża liczba obserwacji ma także kluczowe znaczenie w zastosowaniach centralnego twierdzenia granicznego, pozwalającego na czerpanie wartościowej wiedzy z realizacji ciągu niezależnych zmiennych losowych o identycznym rozkładzie, bez znajomości konkretnej postaci analitycznej tego rozkładu (skokowego lub ciągłego). Duże próby umożliwiają także wykorzystanie estymatorów asymptotycznie nieobciążonych i asymptotycznie efektywnych, redukcję dyspersji estymatorów, zmniejszenie rozpiętości przedziałów ufności czy poprawę mocy testów statystycznych.

O ile korzyści związane z dużą liczbą obserwacji w próbach były ewidentne, o tyle przez długi czas łatwo było przeoczyć potencjalne zagrożenia. Przede wszystkim dlatego, że gromadzenie i przetwarzanie dużych i bardzo dużych zbiorów danych było w przeszłości trudne. Dodatkowo uwaga badaczy i praktyków stosujących metody wnioskowania statystycznego skupiała się niemal wyłącznie na jednym z kilku

[1] Warto zwrócić uwagę, że taka zależność nie jest oczywista i nie musi zachodzić w sytuacjach zastosowania klasycznej lub subiektywnej (personalistycznej) interpretacji prawdopodobieństwa. Szerzej na ten temat pisze m.in. Szreder (2017).

rodzajów błędów, jakie mają prawo się pojawić we wnioskowaniu. Chodzi tu o błąd losowy, a ściślej – błąd losowania (ang. *sampling error*), który – zdaniem Lesliego Kisha (Platek i Särndal, 2001) – był przez naukowców „nadmiernie badany"[2]. Tym stwierdzeniem chciał zapewne zwrócić uwagę na pomijanie przez statystyków innych, nielosowych elementów całkowitego błędu badania próbkowego (ang. *total survey error*). Te ostatnie zaś, w przeciwieństwie do błędu losowania, nie muszą się zmniejszać wraz ze wzrostem liczebności próby. Tymczasem wielu użytkowników metod statystycznych, w szczególności w badaniach społecznych – nie zdając sobie sprawy z natury takich błędów, jak: błąd pokrycia (ang. *coverage error*), błędy braków odpowiedzi (ang. *non-response errors*), błędy pomiaru (ang. *measurement errors*) czy błędy przetwarzania danych (ang. *data processing errors*) – zakłada implicite, że ich negatywny wpływ na jakość wnioskowania zmniejsza się wraz ze wzrostem wielkości próby. W rzeczywistości taka zależność nie istnieje, przede wszystkim dlatego, że błędy nielosowe najczęściej mają charakter systematyczny, tzn. generują obciążenie niebędące funkcją liczebności próby.

Współcześnie dotarcie do różnych grup potencjalnych respondentów, w tym do internetowych paneli respondentów, nie nastręcza trudności; podobną łatwością charakteryzują się możliwości gromadzenia, przetwarzania i analizowania dużych zbiorów danych, co stanowi źródło nowych pokus i iluzji. Do największych pokus należy chęć wykorzystania w badaniu statystycznym nowoczesnego, szeroko dostępnego oprogramowania statystycznego, bez względu na założenia, jakie w odniesieniu do danych statystycznych powinny zostać spełnione w celu zapewnienia poprawności interpretacyjnej wyników. Iluzoryczne zaś jest przekonanie o tym, że skutki ewentualnego niespełnienia pewnych założeń będą niwelowane przez odpowiednio dużą liczbę obserwacji w próbie. W zastosowaniach wnioskowania statystycznego coraz częściej nie zostaje spełnione fundamentalne założenie o posiadaniu przez badacza próby losowej (probabilistycznej).

Celem niniejszego artykułu jest wskazanie i omówienie konsekwencji dostrzegania w dużych próbach statystycznych jedynie szans, a pomijanie wyzwań i zagrożeń wynikających z ich stosowania.

[2] „Sampling error is »over-researched«". Sformułowanie to pojawiło się w artykule Platka i Särndala *Can a statistician deliver?*, opublikowanym w języku polskim wraz z dyskusją w nr. 4/2001 „Wiadomości Statystycznych" (Platek i Särndal, 2001). Inny znany statystyk Prasanta Chandra Mahalanobis zwracał uwagę na ten problem już w 1951 r. Pisał: „In fact, the general attitude is to look upon the non-sampling error as something, which does not concern the statistician, or in any case is a kind of dirty job, which a highbrow statistician need not bother about" (Mahalanobis, 1951, s. 4; „Ogólna postawa jest taka, że błąd nielosowy postrzega się jako coś, co nie dotyczy statystyka, a w każdym razie uważany jest za niezbyt jasną kwestię, którą dumny statystyk nie musi się przejmować". Tłumaczenie tego i pozostałych cytatów w artykule – Mirosław Szreder).

2. Próby nielosowe i ich wielkość a możliwości wnioskowania statystycznego

W wielu badaniach z zakresu problematyki społecznej i ekonomicznej obserwuje się częstsze niż w przeszłości wykorzystanie metod statystycznych, w szczególności wnioskowania statystycznego. Część autorów tego rodzaju badań koncentruje się na zastosowaniu różnych metod estymacji i weryfikacji hipotez, przywiązując niewielką wagę do wymogów związanych z charakterem próby badawczej będącej podstawą wnioskowania. Badacze stosujący nowoczesne i coraz bardziej zaawansowane oprogramowanie statystyczne oczekują przede wszystkim uzyskania wyników istotnych statystycznie, pozwalających na wnioskowanie o populacji, którą reprezentuje próba. Istotność statystyczna stała się w ostatnich kilkunastu latach nie tylko najbardziej pożądanym celem badań próbkowych, lecz także kategorią coraz częściej traktowaną jako swego rodzaju fetysz. Nie może więc dziwić, że statystycy dostrzegający zagrożenia z tym związane zaczęli się sprzeciwiać dotychczasowemu, niekiedy bezkrytycznemu przywiązaniu do istotności statystycznej. Wyrazem tego sprzeciwu były m.in. komentarz redakcyjny w „Nature" z 2019 r., zatytułowany *Scientists rise up against statistical significance*[3] (Amrhein i in., 2019, s. 305) czy opracowanie Gelmana i Sterna (2006) pt. *The Difference Between "Significant" and "Not Significant" is not Itself Statistically Significant*[4], a także szeroko dyskutowane na świecie oświadczenie Amerykańskiego Towarzystwa Statystycznego z 2016 r. na temat istotności statystycznej i rozstrzygania o niej za pomocą wskaźnika *p-value* (Wasserstein i Lazar, 2016)[5].

Wydaje się, że jednym ze słabo do tej pory akcentowanych źródeł tego poruszenia w środowisku naukowym jest dostrzeżenie, że w pogoni za istotnością statystyczną i chęcią wykorzystania nowoczesnego, zwykle przyjaznego oprogramowania statystycznego za mało ważne uznano spełnienie wszystkich założeń matematycznego modelu wnioskowania. W szczególności dotyczy to założenia o obserwacjach w próbie, które – jak wiadomo – powinny być generowane przez mechanizm losowy. Tylko wówczas bowiem, gdy możliwe jest określenie prawdopodobieństwa uzyskania w próbie dowolnego podzbioru jednostek populacji, próbę traktować można jako losową (zob. Barnett, 1991, s. 16; Bracha, 1998, s. 18; Szreder, 2010, s. 68).

Możliwość wykorzystania mediów społecznościowych, elektronicznych list wysyłkowych oraz innych narzędzi elektronicznego kontaktu z przedstawicielami różnych zbiorowości stanowi dla wielu badaczy dużą zachętę do realizacji badań niewyczerpujących, opartych na próbach nielosowych. Coraz powszechniejsze w tego typu

[3] „Naukowcy jednoczą się przeciw istotności statystycznej".
[4] „Różnica między »istotnością« a »nieistotnością« sama w sobie nie jest statystycznie istotna".
[5] Szerzej na ten temat zob. m.in. Szreder (2019). Warto dodać, że już w 2018 r. grupa 72 naukowców zaapelowała na łamach „Nature Human Behaviour" o nowe zdefiniowanie istotności statystycznej (Benjamin i in., 2018).

badaniach jest pozyskiwanie respondentów na zasadzie autoselekcji, czyli dobrowolnego zgłoszenia się do udziału w badaniu, np. poprzez wypełnienie ankiety na stronie internetowej (ang. *opt-in online surveys* lub *self-selection surveys*). Warto więc zwrócić uwagę na konsekwencje takiego doboru jednostek do próby na etapie wnioskowania[6].

Po pierwsze w nielosowym doborze respondentów, który najczęściej wiąże się z brakiem dobrej jakości operatu losowania, istnieje spore ryzyko powstania błędu pokrycia. Brak reprezentantów części populacji w badanej próbie może skutkować obciążeniem estymatorów, którego wielkość i kierunek zwykle są trudne do określenia. W tego typu sytuacjach zwiększanie liczebności próby nie eliminuje – ani nawet nie zmniejsza – powstałego obciążenia. Hirschauer i in. (2020) podkreślają ponadto, że nawet gdyby w omawianym sposobie rekrutowania respondentów użyć kompletnego operatu i nie wystąpiłby błąd pokrycia, to trudno byłoby bronić tezy, że osoby, które zgłosiły się do panelu respondentów, nie różnią się od tych, które odmówiły.

Głośnym przykładem ilustrującym możliwość wystąpienia tego rodzaju błędu stała się w ostatnich miesiącach krytyka, jaką Francuska Akademia Katolicka (Académie catholique de France) skierowała pod adresem znanego francuskiego instytutu badania opinii publicznej IFOP (Institut français d'opinion publique), prowadzącego badania nadużyć seksualnych we francuskim Kościele[7]. Oceny liczbowe i wnioski zostały oparte na próbkowym badaniu ok. 28 tys. dorosłych osób wybranych do próby techniką wyboru kwotowego (nielosowego; Sauvé, 2021, s. 5)[8]. Mimo że liczebność próby, na której oparto wnioskowanie, była wielokrotnie większa od typowej w badaniach opinii społecznej, za słuszne należy uznać zarzuty co do jej reprezentatywności. IFOP skorzystał bowiem z własnego panelu wolontariuszy – osób przyzwyczajonych do tego typu ankiet. Zasadny jest więc – podnoszony przez krytyków – argument o ich pokoleniowej i kulturowej specyfice.

Źródłem obciążenia próby mogą być takie cechy respondentów – odrębne zwłaszcza od cech osób kultywujących tradycyjne wartości oraz osób starszych – jak znajomość sieci społecznościowych i swoboda poruszania się w nich, posiadanie adresu mailowego czy poziom kompetencji cyfrowych. Rezygnacja z postulatu losowego doboru respondentów do próby dała podstawy do uzasadnionej krytyki wyników (i wniosków z) badania. Twierdzenie przedstawicieli IFOP, że podstawą wnioskowania była reprezentatywna – chociaż nielosowa – próba, trzeba uzupełnić o wyrażenie:

[6] Amerykańskie Stowarzyszenie Badaczy Opinii Publicznej (American Association for Public Opinion Research – AAPOR) poświęciło badaniom opartym na internetowych próbach wolontariuszy specjalny 81-stronicowy dokument (zob. American Association for Public Opinion Research, 2010). Najważniejsze wnioski z tego dokumentu zawarto w aneksie do niniejszego opracowania.

[7] Na podstawie: https://www.catholicnewsagency.com/news/249749/french-catholic-academy-critiques-landmark-abuse-report.

[8] Ogólnokrajowe badanie sondażowe odbyło się pomiędzy 25 listopada 2020 r. a 28 stycznia 2021 r.

reprezentatywna, ale jedynie w odniesieniu do cech populacji, które IFOP uznał za ważne i które był w stanie skontrolować. Kilka z nich nie mieści się natomiast w zbiorze charakterystyk kontrolnych, a mogą istotnie różnicować osoby reprezentowane i niereprezentowane w próbie. Ostateczny efekt wywołany tym zróżnicowaniem jest analogiczny do skutków niekompletnego operatu losowania lub braków odpowiedzi o charakterze nielosowym (ang. *missing data not at random*). Często jest to efekt mający postać systematycznego obciążenia wyników próby, którego nie zmniejsza wzrost jej liczebności. „Duże *n* nic nie znaczy, gdy próbkowanie jest obciążone"[9] – podkreśla były prezes Amerykańskiego Towarzystwa Statystycznego (Cochran, 2015, s. 17). Na błąd próbkowania składa się bowiem zarówno element losowy, będący funkcją liczebności próby, jak i składnik nielosowy, o charakterze systematycznym, na który nie wpływa wielkość próby.

Po drugie wykorzystanie we wnioskowaniu próby nielosowej powoduje, że badacz nie ma wystarczających podstaw do wyznaczenia rozkładów prawdopodobieństwa statystyk z próby. Oderwanie schematu doboru jednostek do próby od mechanizmu losującego sprawia, że traci sens częstościowa interpretacja prawdopodobieństwa, będąca podstawą wyznaczania rozkładów wszelkich statystyk, a także przypisywania określonych właściwości estymatorom i testom statystycznym. W rezultacie nie jest możliwe odnoszenie interpretacji probabilistycznej ani do przedziałów ufności, ani do błędów pierwszego i drugiego rodzaju w testowaniu hipotez, ani też do popularnego w oprogramowaniu statystycznym wskaźnika *p-value*.

Nie oznacza to jednak, że próby nieprobabilistyczne są bezużyteczne w poszukiwaniach wzorców i prawidłowości odnoszących się do populacji, z których pochodzą. Specyfika wyboru tych prób sprawia, że przed ich ewentualnym – i warto podkreślić: ostrożnym – wykorzystaniem do wnioskowania powinny podlegać skrupulatnej analizie pod kątem stopnia ich reprezentatywności. Należy w tym miejscu zauważyć, że nie ma ścisłej matematycznej definicji *reprezentatywności*, można jednak wywieść jej istotę np. z twierdzenia Bernoulliego (zob. np. Kordos, 2014). Celem takiej analizy powinno być zatem skonfrontowanie struktury otrzymanej próby z innymi źródłami danych o populacji, którą próba ta reprezentuje. I co ważne – alternatywą dla tego typu działań wzbogacających informację próbkową nie może być jedynie zwiększanie *n*, czyli liczby obserwacji, ponieważ wraz ze wzrostem liczebności próby maleje jedynie składnik losowy błędu próbkowania, ale nie zmniejsza się składnik systematyczny (błąd o charakterze nielosowym).

Duża lub bardzo duża próba nie jest właściwą odpowiedzią na niemożność wykorzystania mechanizmu losującego w próbkowaniu. Jednym z najczęściej współcześnie stosowanych podejść do niwelowania błędu wyboru (ang. *selection bias*) jest

[9] „A large *n* means nothing if the sampling is biased".

wykorzystanie modeli *propensity score*, których istota w badaniach niewyczerpujących sprowadza się do wykorzystania informacji o rozkładach zmiennych ważnych ze względu na cel badania, służących do oszacowania indywidualnego prawdopodobieństwa lub skłonności wzięcia przez respondenta udziału w badaniu vs odmowy udziału w badaniu[10]. Ogólniej – wykorzystanie informacji spoza próby, m.in. w technikach ważenia, kalibracji czy parowania (ang. *sample matching*)[11], albo informacji o innych zmiennych (np. w modelu regresyjnym) może przyczynić się do zwiększenia stopnia reprezentatywności próby. Ale próba nielosowa, nawet wzbogacona o dodatkowe informacje, nie może być uważana za odpowiadającą wymogom matematycznego modelu wnioskowania. Odwołanie się zaś do modelu, a nie jedynie do rzeczywistości, którą model opisuje, jest ważne, ponieważ – jak słusznie zauważają Amrhein i in. (2019, s. 262) – wnioskowanie statystyczne jest myślowym eksperymentem opisującym przewidywaną reakcję modelu na dokonane obserwacje rzeczywistości[12].

Warte podkreślenia jest to, że zarówno teoria statystyki matematycznej, jak i praktyka badań niewyczerpujących dostarczają wielu argumentów na rzecz stosowania we wnioskowaniu jedynie prób probabilistycznych, dla których alternatywą nie może być próba nielosowa – nawet wówczas, gdy jej liczebność jest duża. Nie należy tego stwierdzenia traktować jako oczywistego i lekceważyć obserwowanej w ostatnich latach presji części badaczy, aby złagodzić wymogi odnoszące się do sposobów uzyskiwania obserwacji w próbie.

W Polsce rzadziej niż w innych krajach dyskutuje się o tym problemie, ponieważ prawdopodobnie zakłada się, że skoro dla statystyków jest on zrozumiały i oczywisty, to powinien być taki również dla innych użytkowników metod statystycznych. W rzeczywistości niedocenianie znaczenia założenia o losowości próby wydaje się jednym z powodów kryzysu replikowalności w naukach społecznych i przyrodniczych, a także jedną z ważnych przyczyn malejącego zaufania do badań sondażowych opinii społecznej, a może i do badań statystycznych w ogóle. Dlatego m.in. Hirschauer i in. (2020, s. 84) przypomnieli i uzasadnili, że w przypadkach prób nieprobabilistycznych, których nie udało się zmodyfikować na podstawie rzetelnych informacji o populacji, badacz powinien powstrzymać się od sugerowania, że struktura próby jest nieobciążona[13]. Rok później ci sami autorzy powtórzyli: „Ogólnie rzecz biorąc, wszystkie procedury wnioskowania statystycznego oparte na błędzie standardowym – łącznie z testami istotności – są nieodpowiednie do tego, aby wnioskować

[10] Podejście to wprowadzili do statystyki Rosenbaum i Rubin (1983). Szerzej o tych modelach piszą m.in. Guo i Fraser (2015), Hirschauer i in. (2020), a także Mercer i in. (2017).

[11] Zob. np. Kozłowski i Szreder (2020), Särndal i Lundström (2006) lub Szymkowiak (2019).

[12] „Statistical inference is a thought experiment, describing the predictive performance of models about reality".

[13] „ […] we should refrain from delusively insinuating that the composition of the sample is unbiased".

o populacji na podstawie próby, gdy badanie jest oparte na próbie nieprobabilistycznej”[14] (Hirschauer i in., 2021, s. 21).

W praktyce teoria ta znajduje potwierdzenie w wielu badaniach prowadzonych w różnych okolicznościach. Jednym z przykładów mogą być wnioski wysnute z naukowych analiz źródeł błędów w sondażach przed wyborami parlamentarnymi w Wielkiej Brytanii w 2015 r. Zespół badaczy pod kierunkiem prof. Patricka Sturgisa, powołany przez Royal Statistical Society do zbadania przyczyn niepowodzeń badań sondażowych, jedną z rekomendacji zawartych w swoim raporcie sformułował następująco: „Instytuty badawcze powinny podjąć działania pozwalające na uzyskiwanie bardziej reprezentatywnych prób w obrębie ważonych komórek, które zostały określone”[15] (Sturgis i in., 2016, s. 72). W objaśnieniach do tego postulatu autorzy dodają: „Chętnie widzielibyśmy wykorzystanie w sondażach prawdziwie losowych schematów wyboru próby”[16] (Sturgis i in., 2016, s. 72).

Autor innego opracowania na ten temat, zatytułowanego *Korzyści z próbkowania losowego. Lekcje z brytyjskich wyborów parlamentarnych 2015*[17] (Curtice, 2016), podkreśla, że próbkowanie losowe jest wciąż dużo pewniejszym sposobem zapewnienia reprezentatywności próby, a przez to solidniejszym fundamentem wnioskowania o populacji w porównaniu z innymi technikami próbkowania. Do analogicznych wniosków, wskazujących na niewłaściwe mechanizmy próbkowania i ważenia jako w największym stopniu odpowiedzialne za nieudane sondaże w brytyjskich wyborach w 2015 r., doszli Mellon i Prosser (2017, s. 683): „[...] problemy z próbkowaniem i ważeniem odegrały znaczącą rolę w całym tym sondażowym zamieszaniu”[18]. Wydaje się więc, że warunkiem poprawy jakości sondaży jest większy udział prób probabilistycznych albo doskonalenie modeli korekty struktur prób nielosowych na podstawie dodatkowych informacji spoza próby[19].

3. Duża liczebność próby w testowaniu hipotez statystycznych

W Polsce niemała grupa użytkowników metod statystycznych z oporami przyjmuje do wiadomości krytykę testowania hipotez statystycznych (procedury testowania istotności statystycznej hipotezy zerowej, ang. *null hypothesis significance testing procedure* – NHSTP) i kultu istotności (tak zatytułowali swoją książkę Ziliak

[14] „More generally speaking, all statistical inferential procedures based on the standard error – including statistical significance tests – are inappropriate for making sample-to-population inferences when studies are based on non-random samples”.
[15] „Pollsters should take measures to obtain more representative samples within the weighting cells they employ”.
[16] „We would very much welcome the implementation of truly random sample designs […]”.
[17] „The Benefits of Random Sampling. Lessons from the 2015 UK General Election”.
[18] „[…] problems with sampling and weighting played a substantial role in the polling miss”.
[19] Szeroko o wykorzystaniu informacji spoza próby w badaniach niewyczerpujących piszą Kozłowski i Szreder (2020).

i McCloskey, 2008), chociaż część badaczy ma świadomość, że wskaźnik *p-value* nie może być jedynym wyznacznikiem istotności – należy także brać pod uwagę moc stosowanego testu, jego rozmiar czy inne czynniki mające wpływ na istotność (szerzej rozważa to np. Szymczak, 2018, wskazując na wszechobecny „kult *p-value*").

Tymczasem w świetle coraz silniej akcentowanej różnicy między istotnością merytoryczną (praktyczną) a statystyczną oraz krytyki rozstrzygania o tej ostatniej na podstawie jedynie wskaźnika *p-value* wiele redakcji czasopism zdążyło już zmienić zasady publikowania tekstów zawierających określone elementy wnioskowania statystycznego. Najbardziej radykalny pod tym względem okazał się periodyk „Basic and Applied Social Psychology", którego redakcja już w 2014 r. poinformowała, że nie będzie akceptować prac wykorzystujących procedurę NHSTP z powodu jej ułomności[20] (Trafimow, 2014, s. 1). Ciekawą dyskusję na temat efektywności wprowadzenia tego ograniczenia przeprowadzili Fricker Jr. i in. (2019). Niemniej wśród najważniejszych źródeł powstałego problemu rzadko wymienia się liczebność próby i rozumienie jej roli w testach statystycznych, chociaż jest ona, co postaramy się niżej uzasadnić, jedną z ważnych przyczyn kryzysu istotności statystycznej.

Nie przypadkiem słabości statystycznej weryfikacji hipotez i kryzys replikowalności doświadczeń ujawniły się z dużą siłą dopiero współcześnie, gdy badacze zyskali wielkie, nieporównywalne z dawniejszymi możliwości pobierania dużych prób i analizowania dużych zbiorów danych za pomocą szybkiego i efektywnego oprogramowania komputerowego. W tych warunkach znacznie łatwiejsze stało się wykazanie istnienia statystycznie istotnych różnic na podstawie sprawnie obliczonego wskaźnika *p-value*. Przez długi czas wykazanie istotności statystycznej – często będące warunkiem publikacji artykułu naukowego – wydawało się ważniejsze od uzasadnienia merytorycznej istotności zaobserwowanych różnic. Niezależnie od dziedziny zastosowań wnioskowania statystycznego uzyskanie wyników statystycznie istotnych było stosunkowo łatwe. Często sięgano po prostu do coraz większych liczebnie prób albo decydowano się na wielokrotne próbkowanie, aż do momentu uzyskania satysfakcjonującego rezultatu. Pierwsze z tych postępowań ma ważny wymiar merytoryczny, a drugie – przede wszystkim etyczny.

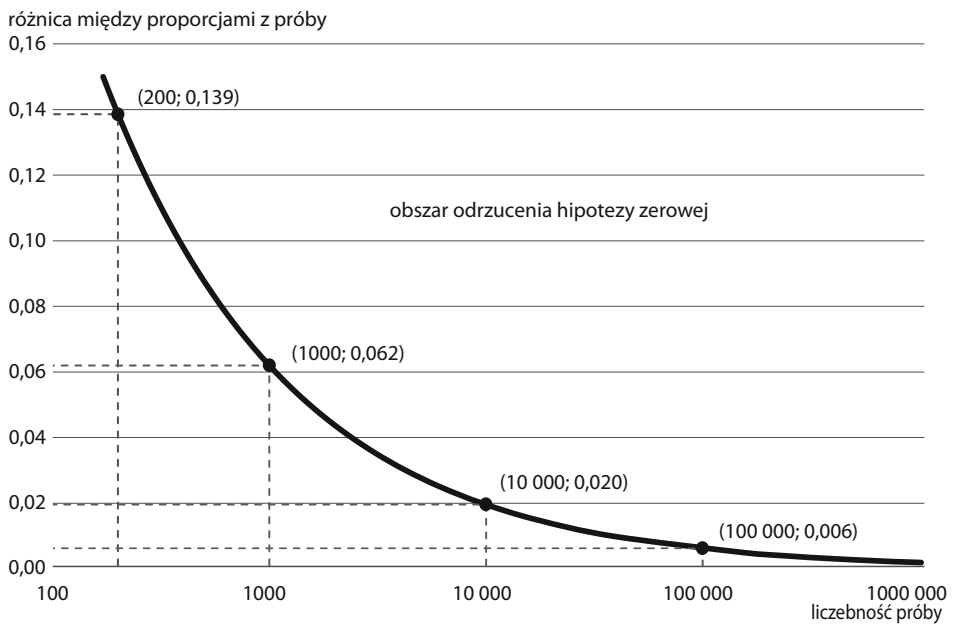
Gdy zwiększamy liczebność próby, rosną szanse na odrzucenie hipotezy zerowej o braku efektu – nawet wówczas, gdy zaobserwowany w próbie efekt jest mały i merytorycznie bez znaczenia.

Przykład. Załóżmy, że weryfikacji podlega hipoteza o jednakowej proporcji palaczy wśród kobiet (p_1) i mężczyzn (p_2):
$H_0: p_1 = p_2$
$H_1: p_1 \neq p_2$

[20] „The null hypothesis significance testing procedure (NHSTP) is invalid [...]" („Procedura testowania istotności statystycznej hipotezy zerowej jest ułomna [...]").

Z wcześniejszych badań wiadomo, że proporcje te są identyczne i wynoszą 0,50. Badacz ma do dyspozycji próbę o liczebności *n*, składającą się w połowie z kobiet i w połowie z mężczyzn. Na wykresie na drugim miejscu w parach liczb ujętych w nawiasach podano, jaka różnica w próbie między proporcją palących kobiet i proporcją palących mężczyzn (wielkość efektu) zostaje uznana za statystycznie istotną, czyli daje podstawę do odrzucenia sprawdzanej (prawdziwej) hipotezy na poziomie istotności 0,05. Różnice te podano w ułamkach (nie w procentach) dla różnych liczebności próby badawczej (*n*) – pierwsza liczba z par zapisanych w nawiasach.

Wykres. Różnice między proporcjami z próby, przy których odrzucona zostaje hipoteza o równości proporcji w populacji na poziomie istotności 0,05 (test dwustronny) w zależności od liczebności próby



Źródło: opracowanie własne.

Z rezultatów przedstawionych na wykresie wynika, że przy odpowiedniej wielkości próby nawet niewiele ponad 0,5 p.proc. różnicy w proporcjach z próby (dla n = 100 000) może stanowić podstawę do odrzucenia hipotezy o równości dwóch proporcji w populacji. Dlatego korzystając z procedur testowania hipotez statystycznych, warto mieć świadomość, że „przy odpowiednio dużej wielkości próby w zasadzie wszystkie zależności próbkowe będą statystycznie istotne”[21] (Lempert, 2009,

[21] „With a large enough N, virtually all associations in a sample will be statistically significant”.

s. 230). Analitycznym uzasadnieniem tej prawidłowości jest formuła (Lin i in., 2013), z której wynika, że wartości *p-value* oparte na zgodnych estymatorach charakteryzują się następującą własnością asymptotyczną odnoszącą się do hipotezy zerowej: $H_0: \beta = 0$

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

Równanie (1) jest prawdziwe dla każdego $\varepsilon > 0$. Oznacza to, że wraz ze wzrostem liczebności próby dowolnie mały efekt zaobserwowany w próbie (ang. *size effect*) tworzy wystarczające podstawy – poprzez malejące do 0 wartości *p-value* – do odrzucenia hipotezy zerowej. Nieznaczący zarówno merytorycznie, jak i praktycznie efekt może w dużych próbach zostać łatwo uznany za statystycznie istotny, ponieważ prawdopodobieństwo jego uzyskania przy założeniu prawdziwości hipotezy zerowej (*p-value*) będzie bliskie 0. Między innymi z tego powodu redakcja „Basic and Applied Social Psychology" zrezygnowała z publikowania artykułów wykorzystujących NHSTP i zaleca autorom szersze stosowanie statystyki opisowej oraz podawanie i ocenianie wielkości efektów w próbie.

Korzystając z większych liczebnie prób, łatwiej jest uzyskać wyniki statystycznie istotne dla bardzo wymagającego poziomu istotności (małych prawdopodobieństw popełnienia błędu pierwszego rodzaju), np. 0,001 lub 0,0001. Błędne jest jednak twierdzenie, że odrzucenie hipotezy zerowej na takim poziomie istotności jest równoznaczne z wykazaniem silniejszych zależności niż w sytuacji, gdyby stwierdzono ich istotność statystyczną na poziomie 0,05 lub 0,01[22]. Taka interpretacja, mimo że intuicyjnie frapująca, jest niepoprawna, właśnie ze względu na zależność, jaka istnieje między wielkością efektu i liczebnością próby z jednej strony a wartością *p-value* z drugiej. Zależność tę wyraża wzór (1). Wydaje się, że jest ona kluczowa dla zrozumienia okoliczności, które wywołały problem czy nawet kryzys związany z używaniem wskaźnika *p-value*. Powszechne wykorzystanie tego wskaźnika w oprogramowaniu statystycznym przyczyniło się do uznania go za podstawową statystykę decydującą o odrzuceniu lub nieodrzuceniu testowanej hipotezy zerowej. Coraz większe możliwości korzystania przez badaczy z dużych prób spowodowały jednak, że stosunkowo łatwe stało się uzyskanie pożądanych wartości *p-value* ($p < 0{,}05$), pozwalających na odrzucenie testowanej hipotezy. Przez długi czas nie zwracano uwagi na niedoskonałości tego miernika (o czym piszą autorzy ponad 40 tekstów w specjalnym numerze „The American Statistician" z marca 2019 r.), a także na to, że jako statystyka z próby wskaźnik ten może być obarczony różnymi błędami o charakterze

[22] Szerzej na ten temat pisze m.in. Lempert (2009).

nielosowym[23]. Skoncentrowanie się wyłącznie na jego wartości próbkowej podczas rozstrzygania o odrzuceniu hipotezy zerowej może sprzyjać lekceważeniu ważnych relacji między wielkością efektu, liczebnością próby i wartością *p-value*.

Wspomniana wcześniej kwestia niezwiązana z dużą liczebnością próby, lecz dotycząca dużej liczby powtarzalnego próbkowania – losowania próby do momentu uzyskania statystycznie istotnego efektu, powinna być rozpatrywana łącznie w dwóch aspektach: merytorycznym i etycznym. Może się zdarzyć, że badacz, prezentując wyniki testowania hipotezy, nie poinformuje o fakcie jej odrzucenia w wyniku wielokrotnego próbkowania. Tego rodzaju brak przejrzystości w zastosowaniu procedury NHSTP wprowadza w błąd odbiorców, ponieważ ukrywa prawdziwe w takim postępowaniu prawdopodobieństwo popełnienia błędu pierwszego rodzaju (α) i naraża badacza na słuszny zarzut nieprzestrzegania zasad etyki w nauce. Jeśli deklarowany poziom istotności (α) wynosi 0,05, to z jego interpretacji wynika, że przeciętnie raz na 20 losowych prób statystyka testowa przyjmie wartość liczbową prowadzącą do odrzucenia hipotezy, która w rzeczywistości jest prawdziwa. Oznacza to, że faktyczny brak efektu w populacji będzie raz na 20 losowań (przeciętnie) upoważniał badacza do odrzucenia hipotezy o braku efektu, a tym samym do stwierdzenia istnienia statystycznie istotnego efektu w populacji. Jeżeli więc ukryje się wyniki statystyk z próby dla 19 losowań i wyeksponuje się jedynie losowanie, w którym wartość statystyki doprowadziła do odrzucenia hipotezy zerowej, to zadeklarowany na wstępie poziom istotności nie będzie równy 0,05 (5%), lecz[24]

$$1 - 0{,}95^{20} \approx 0{,}642 \approx 64{,}2\%. \tag{2}$$

W tych okolicznościach wskazane przez badacza prawdopodobieństwo popełnienia błędu pierwszego rodzaju jest w rzeczywistości ponad 12-krotnie większe i przekracza 60%.

Zasygnalizowany wyżej problem wiąże się z dużą liczbą losowań próby, wykonywanych w nadziei na otrzymanie oczekiwanego rezultatu – odrzucenie hipotezy zerowej. Jednak przejawia pewne wspólne cechy z zagadnieniem liczebności próby. Po pierwsze każde z nich ujawniło się i rozpowszechniło dzięki ułatwieniom, jakich dostarczają statystykom nowoczesne narzędzia i techniki gromadzenia i przetwarzania dużych zbiorów danych, w tym coraz efektywniejsze oprogramowanie statystyczne. Po drugie konsekwencje zarówno dużych prób oraz opartych na ich analizie wskaźników *p-value*, jak i wielokrotnego próbkowania tudzież braku przejrzystości

[23] Gelman (2013, s. 70) ujmuje to następująco: „Syntetycznie wskaźnik *p-value* sam w sobie jest statystyką i może być zaszumioną (niedokładną) miarą przesłanek" („In short, the P value is itself a statistic and can be a noisy measure of evidence").

[24] Oblicza się go łatwo, korzystając z rozkładu dwumianowego i dopełnienia do 1 prawdopodobieństwa, że próba prowadząca do odrzucenia prawdziwej hipotezy nie pojawi się ani razu w 20 losowaniach.

w zakresie informowania o tym przyczyniły się do znacznego zintensyfikowania dyskusji o użyteczności teorii weryfikacji hipotez statystycznych i adekwatności praktycznych procedur rozstrzygania o odrzuceniu lub nieodrzuceniu testowanej hipotezy.

4. Podsumowanie

W artykule podkreślono konieczność interpretowania wyników badania niewyczerpującego w kontekście wielkości próby, która stanowi podstawę wnioskowania. Uzasadniono m.in., że duża liczebność próby nie jest w stanie zrekompensować braku wykorzystania mechanizmu losującego w generowaniu obserwacji próbkowych. Dotyczy to w szczególności coraz powszechniejszych w naukach społecznych badań opartych na internetowych, nieprobabilistycznych panelach respondentów.

Wskazano ponadto, że w przypadku badań reprezentacyjnych należy dopuszczać możliwość wystąpienia w błędzie losowania – oprócz składnika losowego – elementu systematycznego (obciążenia), niebędącego malejącą funkcją liczebności próby. Częstym jego źródłem jest niedoskonałe odzwierciedlenie zbioru jednostek badanej populacji w zastosowanym operacie losowania, czyli wystąpienie błędu pokrycia.

Wielkość próby badawczej ma istotne znaczenie w rozstrzyganiu o odrzuceniu lub nieodrzuceniu sprawdzanej hipotezy statystycznej. W wyniku zwiększania wielkości próby łatwiejsze staje się bowiem odrzucenie hipotezy zerowej, nawet przy mało znaczących (przypadkowych albo spowodowanych błędami nielosowymi) efektach próbkowych. Dlatego w praktyce stosowania procedury testowania hipotez ważne jest, aby decyzje o uznaniu efektu w próbie za statystycznie istotny były poparte wnikliwą analizą, wykraczającą poza ocenę jednego wskaźnika (*p-value*) i powiązaną z liczebnością próby.

Bibliografia

American Association for Public Opinion Research. (2010, czerwiec). *AAPOR Report on Online Panels*. https://www.aapor.org/Education-Resources/Reports/Report-on-Online-Panels.aspx.

Amrhein, V., Greenland, S., McShane, B. (2019, 20 marca). *Scientists rise up against statistical significance*. https://www.nature.com/articles/d41586-019-00857-9.

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.

Barnett, V. (1991). *Sample Survey. Principles and Methods* (2nd edition). Edward Arnold.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Eferson, C., ... Johnson, V. E.

(2018). Redefine statistical significance. *Nature Human Behaviour, 2*(1), 6–10. https://doi.org/10.1038/s41562-017-0189-z.

Bracha, C. (1998). *Metoda reprezentacyjna w badaniu opinii publicznej i marketingu*. Efekt.

Cochran, J. (2015, 1 lutego). *ASA Leaders Reminisce. Peter (Tony) Lachenbruch*. http://magazine.amstat.org/blog/2015/02/01/peterlachenbruch_feb2015/.

Curtice, J. (2016). The Benefits of Random Sampling. Lessons from the 2015 UK General Election. *NatCen Social Research*. https://www.bsa.natcen.ac.uk/media/39018/random-sampling.pdf.

Fricker Jr., R. D., Burke, K., Han, X., Woodall, W. H. (2019). Assessing the Statistical Analyses Used in *Basic and Applied Social Psychology* After Their *p*-Value Ban. *The American Statistician*, *73*(Sup1), 374–384. https://doi.org/10.1080/00031305.2018.1537892.

Gelman, A. (2013). P values and statistical practice. *Epidemiology*, *24*(1), 69–72. https://doi.org/10.1097/EDE.0b013e31827886f7.

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.

Guo, S., Fraser, M. W. (2015). *Propensity Score Analysis. Statistical Methods and Applications* (2nd edition). Sage Publications.

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.

Hirschauer, N., Grüner, S., Mußhoff, O., Becker, C., Jantsch, A. (2021). Inference using non-random samples? Stop right there!. *Significance*, *18*(5), 20–24. https://doi.org/10.1111/1740-9713.01568.

Kordos, J. (2014). Od twierdzenia Jakuba Bernoulliego do współczesnych badań reprezentacyjnych. *Wiadomości Statystyczne*, *59*(3), 1–23. https://ws.stat.gov.pl/Article/2014/3/001-023.

Kozłowski, A., Szreder, M. (2020). *Informacje spoza próby w badaniach statystycznych*. Wydawnictwo Uniwersytetu Gdańskiego.

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

Mahalanobis, P. C. (1951). Professional Training in Statistics. *Bulletin of the International Statistical Institute*, *33*(5), 335–342.

Mellon, J., Prosser, C. (2017). Missing nonvoters and misweighted samples. Explaining the 2015 great British polling miss. *Public Opinion Quarterly*, *81*(3), 661–687. https://doi.org/10.1093/poq/nfx015.

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*(1), 250–271. https://doi.org/10.1093/poq/nfw060.

Platek, R., Särndal, C. E. (2001). Czy statystyk może dostarczyć dane wysokiej jakości?. *Wiadomości Statystyczne*, *46*(4), 1–21.

Rosenbaum, P. R., Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55. https://doi.org/10.1093/biomet/70.1.41.

Sauvé, J.-M. (2021). *Sexual Violence in the Catholic Church France 1950–2020. Summary of the Final Report Independent Commission on Sexual Abuse in the Catholic Church (CIASE)*. https://www.ciase.fr/medias/Ciase-Summary-of-the-Final-Report-5-october-2021.pdf.

Särndal, C. E., Lundström, S. (2006). *Estimation in Surveys with Nonresponse*. John Wiley & Sons.

Sturgis, P., Baker, N., Callegaro, M., Fisher, S., Green, J., Jennings, W., Kuha, J., Lauderdale, B., Smith, P. (2016). *Report of the Inquiry into the 2015 British general election opinion polls*. London: Market Research Society and British Polling Council. https://eprints.ncrm.ac.uk/id/eprint/3789/1/Report_final_revised.pdf.

Szreder, M. (2010). *Metody i techniki sondażowych badań opinii*. Polskie Wydawnictwo Ekonomiczne.

Szreder, M. (2017). Nowe źródła informacji i ich wykorzystanie w podejmowaniu decyzji. *Wiadomości Statystyczne*, *62*(7), 5–17. https://doi.org/10.5604/01.3001.0014.0972.

Szreder, M. (2019). Istotność statystyczna w czasach big data. *Wiadomości Statystyczne. The Polish Statistician*, *64*(11), 42–57. https://doi.org/10.5604/01.3001.0013.7583.

Szymczak, W. (2018). *Praktyka wnioskowania statystycznego*. Wydawnictwo Uniwersytetu Łódzkiego.

Szymkowiak, M. (2019). *Podejście kalibracyjne w badaniach społeczno-ekonomicznych*. Wydawnictwo Uniwersytetu Ekonomicznego w Poznaniu.

Trafimow, D. (2014). Editorial. *Basic and Applied Social Psychology*, *36*(1), 1–2. https://doi.org/10.1080/01973533.2014.865505.

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*. University of Michigan Press.

Aneks

Amerykańskie Stowarzyszenie Badaczy Opinii Publicznej tak wypowiedziało się o badaniach wykorzystujących internetowe próby wolontariuszy (ang. *opt-in online surveys*)[25]:

- badacze powinni omijać nielosowe internetowe panele respondentów w sytuacjach, gdy jednym z celów badawczych jest precyzyjne oszacowanie parametrów populacji. Brak jest powszechnie akceptowanych podstaw teoretycznych, które uzasadniałyby możliwości wnioskowania o całej populacji na podstawie tego rodzaju nieprobabilistycznych prób. Z tego powodu należy unikać stwierdzeń o „reprezentatywności" w przypadku korzystania z takich źródeł próbkowania;

[25] https://www.aapor.org/Education-Resources/Reports/Report-on-Online-Panels.aspx.

- z większości analiz porównujących wyniki uzyskane z nielosowych paneli internetowych z wynikami pochodzącymi z próbkowania probabilistycznego wynika, że istnieją między nimi istotne różnice w odniesieniu do szerokiej gamy postaw i zachowań respondentów;
- są sytuacje, kiedy nielosowe internetowe panele respondentów – ze względu na relatywnie niskie koszty ich wykorzystania i specyficzne sposoby gromadzenia danych za pomocą sieci internetowej – stanowią właściwy wybór. Dotyczy to w szczególności badań, w których nie oczekuje się precyzyjnej estymacji parametrów populacji;
- użytkownicy paneli internetowych powinni być świadomi tego, że istnieją istotne różnice w strukturze i praktyce wykorzystania poszczególnych paneli, co może rzutować na otrzymane wyniki. Badacze powinni uważnie wybierać panele do swoich badań;
- kluczowe jest pełne ujawnienie sposobu uzyskania danych próbkowych. Tylko wówczas możliwe jest bowiem wiarygodne ocenienie jakości wyników badania i możliwości ich replikowania.