Wiadomości Statystyczne. The Polish Statistician, 2019, vol. 64, 11, 42–57
Studia interdyscyplinarne. Wyzwania badawcze / Interdisciplinary studies. Research challenges

Istotność statystyczna w czasach big data

Mirosław Szreder[a]

Streszczenie. Rozwój nowych technologii wpływa zarówno na realizację badań statystycznych, jak i na postrzeganie ich wyników w świetle innych źródeł informacji. W tym kontekście powraca w środowisku naukowym temat roli testowania hipotez statystycznych oraz interpretowania i przedstawiania jego wyników, w tym stosowania kategorii istotności statystycznej oraz wskaźnika *p*-value. Inspiracją do powstania tego opracowania stała się fala dyskusji wokół tego zagadnienia toczących się na forum czasopism „Nature" i „The American Statistician" na początku 2019 r.

Celem artykułu jest ukazanie szans i zagrożeń, jakie big data stwarza dla weryfikacji hipotez i wnioskowania statystycznego, zarówno w ujęciu klasycznym, jak i w podejściu bayesowskim. Autor uzasadnia konieczność zaniechania zbyt daleko posuniętych uproszczeń w realizacji procesu wnioskowania statystycznego oraz prezentowaniu wyników weryfikacji hipotez. Chodzi zarówno o postulat uwzględnienia jakości danych próbkowych, zwłaszcza typu big data, jak i o podawanie pełnej informacji o modelu statystycznym, na podstawie którego przeprowadza się wnioskowanie.

Słowa kluczowe: wnioskowanie statystyczne, testowanie hipotez, istotność statystyczna, wskaźnik *p*-value, big data, podejście bayesowskie

Statistical significance in the era of big data

Abstract. The development of new technologies has affected both the procedures of traditional statistical surveys and the perception of their results in the light of other available sources of information. In this connection, the role of the verification of statistical hypotheses and of the interpretation and presentation of its results, including the use of statistical significance and *p*-value, has recently returned as a frequent topic for discussion among the scientific community. The author was inspired to write this paper by a wave of discussion regarding this matter held at the beginning of 2019 in the *Nature* and *The American Statistician* journals.

The aim of the paper is to present the opportunities provided and challenges posed by the use of big data to the hypothesis verification process and to statistical inference, both in the traditional and Bayesian approaches. The author explains the necessity of discontinuing adopting excessive simplifications while performing statistical inference and presenting the results of the verification of hypotheses. This involves both the postulate to pay greater attention to the quality of sampling data, especially in the case of data originating from big data sets, as well as the postulate to provide full information about the statistical model on the basis of which the inference is being performed.

Keywords: statistical inference, hypothesis testing, statistical significance, *p*-value, big data, Bayesian approach

JEL: C12, C13, C18, D80

[a] Uniwersytet Gdański, Wydział Zarządzania.

Rosnące znaczenie badań ilościowych w różnych dziedzinach nauki, wspieranych dynamicznym rozwojem technologii komputerowych i technik big data, stanowi zachętę do coraz powszechniejszego stosowania wnioskowania statystycznego. Zjawisko to z jednej strony potwierdza użyteczność i żywotność statystyki, a z drugiej rodzi niebezpieczeństwo formułowania fałszywych wniosków w sytuacji niepoprawnego wykorzystania jej technik i metod. Inne niebezpieczeństwo tkwi w niepełnej prezentacji lub zbyt uproszczonej interpretacji wyników analiz statystycznych. Te zagrożenia kumulują się w szczególności we wnioskowaniu statystycznym, w tym w procesie weryfikacji hipotez w warunkach dostępu badacza do różnych źródeł danych, w tym big data.

Celem niniejszego artykułu jest ukazanie szans i zagrożeń, jakie big data stwarza dla weryfikacji hipotez i wnioskowania statystycznego.

WNIOSKOWANIE DEDUKCYJNE I INDUKCYJNE – RYS HISTORYCZNY

Statystyka rodziła się najpierw jako nauka, która z ludzkiej obserwacji rzeczywistości, powtarzalnych faktów oraz odwzorowań ich częstości w liczbach próbowała wydobyć wiedzę użyteczną dla człowieka. Pierwszymi i ważnymi osiągnięciami statystyków były metody służące do wyodrębniania i charakteryzowania prawidłowości występujących w masowych zjawiskach z różnych dziedzin życia. Współcześnie nazywamy je opisowymi albo metodami statystyki opisowej (ang. *descriptive statistics*). Niemal równolegle, z udziałem filozofów i matematyków, postępował rozwój drugiej ważnej dziedziny statystyki, mianowicie wnioskowania statystycznego (ang. *statistical inference*, *inferential reasoning*). Przełomowym osiągnięciem na tym polu było sformułowanie przez Thomasa Bayesa (1702–1761) twierdzenia pozwalającego określić, jakie przesłanki stojące za empirycznie zaobserwowanymi faktami są najbardziej prawdopodobne. Twierdzenie Bayesa, znane także jako twierdzenie o prawdopodobieństwie przyczyny, stało się podstawą rozwoju indukcyjnego wnioskowania w statystyce[1].

We wnioskowaniu indukcyjnym szacuje się, na podstawie dokonanych obserwacji (próby statystycznej), jaka hipoteza odnosząca się do populacji, z której pochodzi próba, jest najbardziej uzasadniona. W przeciwieństwie do wnioskowania dedukcyjnego – w którym pierwotna jest pewna hipoteza, a zatem określa się, jakie obserwacje w próbie powinny się pojawić w sytuacji, gdyby ta hipoteza była prawdziwa – we wnioskowaniu indukcyjnym punktem wyjścia są obserwacje. Stąd wyniki wnioskowania indukcyjnego nie ograniczają się do jednej ściśle określonej hipotezy. Mogą służyć do formułowania nowych, nieznanych hipotez, czyli wydobywania z danych statystycznych nowej wiedzy.

[1] Szerzej na temat znaczenia przełomowego dla rozwoju rachunku prawdopodobieństwa i statystyki matematycznej twierdzenia Bayesa zob. Szreder (2013).

O tym, że ostatecznie większą popularność w teorii, a zwłaszcza w praktyce statystycznej, zdobyło nie indukcyjne podejście oparte na twierdzeniu Bayesa, lecz wnioskowanie dedukcyjne, zdecydowały dwa główne czynniki (por. m.in. Goodman, 1999). Po pierwsze, w podejściu bayesowskim wymaga się, aby jeszcze przed wylosowaniem próby określić prawdopodobieństwa lub rozkłady *a priori*, odnoszące się do parametrów populacji lub hipotez dotyczących charakterystyk populacji. W praktyce jest to dość kłopotliwe, zarówno ze względu na częsty brak wiedzy wstępnej badacza, jak i trudności z jej wyrażeniem w kategoriach probabilistycznych. Po drugie, prawdopodobieństwa *a priori* rzadko stanowią wynik usystematyzowanych losowych obserwacji jakiegoś doświadczenia, dlatego najczęściej są formułowane nie w kategoriach (obiektywnej) częstości względnej, lecz jako prawdopodobieństwa subiektywne (personalistyczne)[2]. Te z kolei budzą opór u tych badaczy, którzy nie godzą się na dopuszczenie elementów subiektywnych w procesie poznania naukowego. Między innymi z tego powodu już w latach 20. i 30. XX w. poszukiwano innego podejścia do wnioskowania, które opierałoby się na częstościowej interpretacji prawdopodobieństwa. Innymi słowy, chodziło o takie (dedukcyjne) ujęcie procesu wnioskowania, aby przy założeniu prawdziwości sformułowanej na wstępie hipotezy wskazać, z jaką częstością względną w próbie losowej, w warunkach wielokrotnie powtarzanego hipotetycznie losowania, powinny pojawiać się określone wyniki.

Takie częstościowe ujęcie zostało zawarte zarówno w propozycji *p*-value Ronalda A. Fishera z lat 20. XX w., jak i w odrębnej propozycji z wczesnych lat 30. Jerzego Neymana i Egona Pearsona. Wskaźnik *p*-value miał być, według koncepcji Fishera, miernikiem siły przesłanek w próbie, służącym do sfalsyfikowania testowanej hipotezy (hipotezy zerowej). Natomiast Neyman i Pearson oparli swoją koncepcję na dwóch konkurencyjnych hipotezach – zerowej i alternatywnej – i przedstawieniu sposobu podjęcia decyzji (testowania) o odrzuceniu hipotezy zerowej na rzecz alternatywnej lub uznaniu przesłanek (obserwacji) zawartych w próbie za niewystarczające do odrzucenia hipotezy zerowej. Są to dwie różne koncepcje weryfikacji hipotez statystycznych, choć wielu użytkowników metod statystycznych uważa je za elementy jednego wspólnego podejścia do wnioskowania. Szczególną popularność w ostatnich kilkunastu latach zyskała w statystyce i ekonometrii koncepcja *p*-value Fishera, znacznie wygodniejsza dla użytkowników oprogramowania komputerowego od koncepcji testowania Neymana i Pearsona. Programy statystyczne bowiem, wraz z wyliczeniem wartości odpowiedniej statystyki z próby, podają jednocześnie prawdopodobieństwo uzyskania tej właśnie lub mniejszej/większej jej wartości, przy założeniu, że hipoteza zerowa jest prawdzi-

[2] Przez prawdopodobieństwo subiektywne rozumie się stopień przekonania (ang. *degree of belief*) badacza o prawdziwości danego sądu. Twórcami tej interpretacji prawdopodobieństwa byli Ramsey (1926), de Finetti (1937) oraz Savage (1954). O tej i innych interpretacjach prawdopodobieństwa pisze szerzej Szreder (2004).

wa. Prawdopodobieństwo to nazywane jest prawdopodobieństwem krytycznym lub wartością *p*, a częściej z języka angielskiego *p*-value[3]. Jest ono kluczowe w dalszym rozstrzyganiu o losach hipotezy zerowej.

W badaniach statystycznych z różnych dziedzin przyjęło się uważać, że wartość *p* mniejsza od 0,05 świadczy o statystycznej istotności różnicy pomiędzy tym, co zaobserwowano w próbie, a tym, co powinno było wystąpić w próbie, gdyby hipoteza zerowa była prawdziwa. Wynik taki staje się więc podstawą do odrzucenia hipotezy zerowej. Innymi słowy, przyjmuje się, że próg 0,05 jest dla *p*-value rozstrzygający. I mimo że nie wziął się on znikąd, bo zaproponował go sam Fisher[4], to obecnie coraz większa liczba badaczy proponuje odejście od tego progu, a redakcja „The American Statistician" tytułuje cykl ponad 40 artykułów poświęconych współczesnemu testowaniu hipotez *Moving to a World Beyond "p < 0.05"* (Wkraczając do świata poza „p < 0,05").

WSKAŹNIK *P*-VALUE I ISTOTNOŚĆ STATYSTYCZNA – WSPÓŁCZESNA KRYTYKA

Nie jest z pewnością przypadkiem, że czasopismo „The American Statistician" poświęciło w całości swój marcowy numer z 2019 r. krytycznemu omówieniu praktyki wnioskowania statystycznego wykorzystującego koncepcje *p*-value i statystycznej istotności, a równocześnie jedno z najbardziej prestiżowych czasopism naukowych na świecie „Nature" zamieściło obszerny komentarz pt. *Porzućcie przestarzałą istotność statystyczną*. Podjęty w tych czasopismach problem wykracza bowiem znacznie poza narastającą krytykę dychotomizacji wielkości *p*-value w badaniach statystycznych i ekonometrycznych. O wadze poruszanego problemu niech świadczy to, że odnosi się on do samego pojęcia istotności statystycznej – jej oceniania, interpretowania i komunikowania.

Za warte uwagi trzeba uznać przede wszystkim wyrażane przez wielu badaczy wątpliwości co do zasadności posługiwania się jednym dychotomicznym kryterium ($p < 0,05$ lub $p > 0,05$) w rozstrzyganiu o tym, czy coś uznaje się za statystycznie istotne, czy nie. Poleganie na tym prostym rozstrzygnięciu – jak podkreśla m.in. Goodman (1999) – pozbawiło nas niemal zupełnie możliwości rozróżnienia pomiędzy statystycznymi wynikami a konkluzjami naukowymi.

[3] Co do definicji *p*-value panuje zgoda wśród statystyków i badaczy innych dziedzin. Amerykańskie Towarzystwo Statystyczne określa *p*-value jako „prawdopodobieństwo tego, że w warunkach szczegółowo określonego modelu wnioskowania wartość statystyki z próby (miary syntetycznej wyników próby) będzie równa lub przyjmie bardziej ekstremalne wartości od zaobserwowanej" (Wasserstein i Lazar, 2016, s. 129). Tłumaczenie tego i pozostałych cytatów w artykule – Mirosław Szreder.

[4] „The value for which $p = 0.05$ is 1.96 or nearly 2; it is convenient to take this point as a limit in judging whether a deviation is to be considered significant or not" („Wartość, dla której $p = 0,05$, wynosi 1,96 lub prawie 2, stąd wygodne może być przyjęcie tego punktu jako granicznego w ocenie, czy zaobserwowane odchylenie w próbie zostanie uznane za istotne, czy nie") (Fisher, 1925, s. 45).

Tymczasem wyzwaniem dla statystyków jest wciąż to, w jaki sposób wnioski z pojedynczego badania statystycznego wpisać w dłuższy łańcuch kumulowania wiedzy. W przypadku komunikowania wyników z badań warto zaś zdać sobie sprawę z tego, że termin *istotny* jest dla wszystkich tych, którzy nie mają dobrego przygotowania statystycznego, synonimem słów *ważny* lub *znaczący*[5]. Nie zawsze jednak wynik statystycznie istotny oznacza, że jest on ważny. Jeżeli badanie statystyczne wskazuje np., że zastosowanie nowego leku obniża ciśnienie krwi średnio o 0,10 jednostek, z błędem standardowym wynoszącym 0,03 jednostki, to oczywiście wynik taki będzie statystycznie istotny. W praktyce natomiast jego znaczenie będzie niewielkie, biorąc pod uwagę to, że typowy poziom ciśnienia krwi u człowieka wynosi ok. 100 jednostek. I odwrotnie: wynik badania, w którym inny lek obniża ciśnienie krwi o 10 jednostek z błędem standardowym 10 jednostek, będzie statystycznie nieistotny, ale wskazujący na znaczny praktyczny potencjał tego leku w dalszych badaniach[6].

Nie ma ostrej i jednoznacznej linii podziału pomiędzy wyrażeniami *statystycznie istotny* i *statystycznie nieistotny*. Istotność w statystyce zmienia się w sposób ciągły, tak jak ciągłą zmienną losową jest *p*-value. Jedynie dla wygody i uproszczenia jest ona często traktowana dychotomicznie – zero-jedynkowo, co budzi coraz powszechniejszy sprzeciw[7]. Przy dużej liczbie danych w próbie – o co nietrudno w epoce big data – statystyczną istotność da się osiągnąć dosyć łatwo. Tak jak łatwo zwieść mogą badacza korelacje pozorne (ang. *spurious correlations*) i regresje pozorne (ang. *spurious regressions)*. Dlatego warto rozważyć, czy współcześnie *p*-value nie powinno być traktowane tylko jako jeden ze sposobów dowodzenia nieprawdziwości hipotezy zerowej – sposób niewystarczający. Tym bardziej że to prawdopodobieństwo odnosi się nie tylko do hipotezy zerowej, jak przyjęło się uważać. Odnosi się także do całego modelu wnioskowania i jego założeń, a więc także do tych wszystkich okoliczności i zakłóceń (błędów nielosowych), które miały wpływ na przyjęcie w próbie takiej, a nie innej wartości statystyki testowej.

Określona wartość *p* jest zwykle wynikiem działania kombinacji czynników losowych (ang. *random variation*), a także czynników wyrażających konsekwencje odstępstw od przyjętych założeń modelu wnioskowania. Mała *p*-value może oznaczać nieprawdziwość hipotezy zerowej, ale także to, że źle określono niektóre matematyczne założenia modelu albo że obserwacje nie były w pełni generowane przez proces losowy lub też że dały o sobie znać błędy nielosowe[8].

[5] Piszą o tym m.in. Hurlbert, Levine i Utts (2019, s. 354), stwierdzając, że w języku angielskim synonimami terminu *significant* są dla osób spoza środowiska statystyków terminy *important* oraz *influential*.

[6] Przykład zaczerpnięty z: Gelman i Stern (2006, s. 2).

[7] Popularny staje się wyrażany w literaturze anglojęzycznej postulat „*p*-values should not be thresholded" („wartości *p* nie powinny mieć żadnego progu").

[8] Na znaczenie błędów nielosowych wskazują m.in. Stefanowicz i Cierpiał-Wolan (2015) oraz Szreder (2015).

Natomiast duża (większa od 0,05) *p*-value nie dowodzi w żadnym stopniu prawdziwości hipotezy zerowej. Informuje jedynie o stopniu zgodności wyników próby z wartością parametru populacji przyjętą w hipotezie zerowej. W odpowiednim przedziale ufności mieści się bowiem wiele innych wartości tego parametru, z którymi obserwacje próbkowe mogą być zgodne[9].

Świadomość tego, że formułowanie wniosków z badań statystycznych jest ograniczone do pewnego modelu, każe w ich interpretacji uwzględnić cały kontekst relacji model – rzeczywistość, w szczególności założenia poczynione przez badacza. Wnioskowanie statystyczne – jak słusznie przypominają Amrhein, Trafimow i Greenland (2019, s. 262) – jest swego rodzaju eksperymentem myślowym, charakteryzującym przewidywanie reakcji modelu na dokonane obserwacje rzeczywistości. Wyniki wnioskowania odnoszone bezpośrednio do złożonej rzeczywistości zamiast do modelu będącego jej matematycznym uproszczeniem mogą być mylące. Tak jak myląca może być ostrość rozróżnienia między statystyczną istotnością i nieistotnością.

Gelman i Stern (2006) podają przekonujące przykłady na to, że „różnica między istotnością a nieistotnością sama w sobie nie jest statystycznie istotna”[10]. Załóżmy, że dla przetestowania istotności efektu działania pewnego czynnika wykonano dwa badania reprezentacyjne. Z pierwszego otrzymano średnią wielkość efektu równą 25 i odchylenie standardowe równe 10, a z drugiego – średnią wielkość efektu równą 10 i odchylenie standardowe równe 10. Efekt działania czynnika w pierwszym badaniu jest statystycznie istotny, gdyż efekt zerowy (brak efektu) jest oddalony od wartości 25 aż o dwa i pół odchylenia standardowego[11]. W drugim badaniu efekt jest statystycznie nieistotny, bo oddalony jest zaledwie o jedno odchylenie standardowe od zera. Można więc domniemywać, że między wynikami tych dwóch badań istnieje duża różnica. W rzeczywistości różnica ta – obliczona jako wartość oczekiwana różnicy między średnimi próbkowymi – jest statystycznie nieistotna, bo wynosi 15, z odchyleniem standardowym (pierwiastek z sumy kwadratów odchyleń standardowych z obu badań, wynoszącej 200) równym w przybliżeniu 14. Różnica o wartości 0 wykracza bardzo niewiele poza jedno tylko odchylenie standardowe mierzone wokół różnicy uzyskanej w obu badaniach.

Gdyby przeprowadzono trzecie badanie, o znacznie większej liczebności próby, i gdyby dało ono średni efekt równy 2,5, z odchyleniem standardowym 1,0, to ocena efektu zostałaby w nim uznana (analogicznie do pierwszego badania) za statystycznie istotną. Co ciekawe, badania pierwsze i trzecie wskazują na istotne

[9] Szersze wyjaśnienie i przykład można znaleźć m.in. w pracy Szredera (2010).

[10] Wyrażenie ujęte tu w cudzysłów jest tytułem pracy Gelmana i Sterna (2006): *The difference between ”significant” and ”not significant” is not itself statistically significant*.

[11] W rozkładzie normalnym częstość względna obserwacji spoza przedziału: średnia plus/minus dwa i pół odchylenia standardowego wynosi 0,012 (wyraźnie mniej niż 0,05).

oddziaływanie analizowanego czynnika, mimo że istnieją między nimi duże statystycznie istotne różnice. Wartość oczekiwana różnicy średnich wynosi 22,5, z odchyleniem standardowym 10,05[12]. Trzecie badanie potwierdza zatem wynik badania pierwszego, ale tylko w sensie statystycznej istotności oddziaływania analizowanego czynnika. Nie potwierdza natomiast wielkości efektu oddziaływania tego czynnika.

Gelman i Stern (2006) nie są przeciwnikami posługiwania się w testowaniu poziomem istotności, ale stwierdzają, że w takich zagadnieniach, jak zilustrowane powyższym przykładem badacz powinien brać pod uwagę raczej istotność statystyczną różnicy w wynikach z poszczególnych badań, a nie różnicę między poziomami istotności.

Podsumowując, zdaniem autora istota dzisiejszej krytyki posługiwania się wartością *p*, a w dalszej kolejności kategorią istotności statystycznej w testowaniu hipotez, sprowadza się do następującej kwestii: praktyka ograniczania oceny zmiennej losowej ciągłej *p*-value jedynie do rozstrzygnięcia, czy wartość *p* uzyskana w próbie przekracza określone progi (0,05 albo nawet – jak proponują Benjamin i in., 2018 – 0,005), jest niewystarczająca. Trzeba bowiem być świadomym tego, że obliczona *p*-value odnosi się jedynie do wyniku pojedynczej próby oraz że test statystyczny nie jest narzędziem, które może przekształcić niepewność towarzyszącą wnioskowaniu w pewność decyzji dotyczącej prawdziwości lub nieprawdziwości hipotezy. Posługiwanie się tego typu progami powoduje po pierwsze, że część wartościowych i być może dobrze rokujących badań, w których uzyskana wielkość efektu okazała się za mała, aby *p*-value była mniejsza od wartości progowej, zostaje porzucona. Ich wyniki opatruje się zbyt często nieprawdziwymi konkluzjami „nie występuje różnica" lub „nie występuje współzależność" tylko dlatego, że *p*-value okazała się wyższa niż określony próg[13]. Po drugie, w obiegu naukowym sztuczną nadreprezentację uzyskują te prace, w których otrzymano $p < 0,05$, i do ich wyników – jako że innych (statystycznie nieistotnych) redakcje czasopism zwykle nie publikują – badacze przykładają zbyt dużą wagę[14]. Po trzecie zaś, niemal wszyscy uczestnicy dyskusji na temat istotności statystycznej są zgodni co do tego, że we wnioskowaniu statystycznym nie da się wyeliminować niepewności. Jako jeden ze sposobów jej wyrażenia proponuje się rozstrzyganie nie wyłącznie na podstawie *p*-value, ale z uwzględnieniem pełniejszych informacji w procesie weryfikacji hipotez,

[12] Różnica o wartości 0 (brak różnicy) jest zatem oddalona od średniej o ponad dwa odchylenia standardowe.

[13] Podkreślają to wyraźnie autorzy komentarza w „Science" (Amrhein, Greenland i McShane, 2019, s. 305).

[14] Amrhein, Trafimow i Greenland (2019, s. 264) stwierdzają dobitnie: „Nadużywa się statystyki jako maszyny do automatycznego podejmowania naukowych decyzji [*automated scientific decision machine*], zarówno w odniesieniu do weryfikowanych hipotez, jak i w procesie selekcji artykułów kierowanych do publikacji".

w tym o przedziałach ufności[15] dla parametru, którego dotyczy wnioskowanie, lub wartości czynnika Bayesa (ang. *Bayes factor*).

Nieistotność statystyczna może w praktyce oznaczać potrzebę dalszych badań lub konieczność sięgnięcia do innych źródeł informacji mogących pomóc w ocenie zależności ujętej w hipotezie zerowej lub wartości określonego parametru populacji. Wskaźnikowi *p*-value nadano w ostatnich kilkunastu latach zbyt duże znaczenie, sugerujące błędnie, że jest on w stanie wyrazić wszystkie najważniejsze elementy niepewności związane z testowaniem hipotez statystycznych. Obecne dyskusje na ten temat nie zawierają zwykle postulatu rezygnacji z *p*-value, lecz kładą nacisk na potrzebę głębszej analizy źródeł niepewności przed podjęciem decyzji o odrzuceniu lub nieodrzuceniu hipotezy zerowej. Zresztą – jak zauważa Goodman (2019) – trudno byłoby obecnie wycofać się z używania *p*-value, ponieważ wszyscy, w tym naukowcy i redakcje czasopism naukowych, przyzwyczaili się do tego wskaźnika. Umownie przypisuje się mu pewną wartość poznawczą. Zdaniem Goodmana (2019) jest to już zjawisko socjologiczne, nie tylko naukowe. Tak jak wierzymy w wartość pieniądza, oznaczającą w przekonaniu konsumentów prawo do nabycia określonych dóbr lub usług, tak samo przyjmujemy, że za określonymi wartościami *p* kryje się prawo do naukowego uznania określonych hipotez, a często także prawo do opublikowania uzyskanych wyników.

Najlepszym podsumowaniem debaty o istotności statystycznej i *p*-value wydaje się wciąż aktualne, syntetycznie ujęte w sześciu punktach, stanowisko Amerykańskiego Towarzystwa Statystycznego z 2016 r., stanowiące załącznik do niniejszego opracowania.

BIG DATA A PROBLEMY TESTOWANIA HIPOTEZ

Dla wnioskowania statystycznego big data stanowi z jednej strony wyzwanie, gdyż umożliwiając dostęp do znacznie większej liczby obserwacji próbkowych niż w przeszłości, wymaga odrębnego, spójnego podejścia teoretycznego do estymacji i weryfikacji hipotez, jakie do tej pory jeszcze nie powstało. Z drugiej strony zasoby big data rozumiane jako alternatywne źródła informacji lub niestatystyczne źródła danych[16] dają szansę na poprawę jakości wnioskowania w warunkach rosnącej skali i wagi błędów nielosowych, w tym nasilającej się tendencji do odmawiania przez respondentów udziału w badaniach. W kontek-

[15] W komentarzu „Nature" postuluje się, aby przedziały ufności (ang. *confidence intervals*) nazywać raczej przedziałami zgodności z danymi z próby (ang. *compatibility intervals*) oraz analizować i informować odbiorcę wyników, jakie są implikacje tego, że w przedziałach tych mieści się wiele wartości i co oznaczają końce tych przedziałów (Amrhein, Greenland i McShane, 2019, s. 307).

[16] Por. np. Beręsewicz i Szymkowiak (2015).

ście scharakteryzowanej w poprzedniej części artykułu dyskusji na temat istotności statystycznej i *p*-value big data stawia niektóre problemy z zakresu testowania hipotez w ostrzejszym świetle, w przypadku innych zaś może stanowić swoiste remedium.

Pierwszym problemem, który uwidacznia się zwłaszcza w sytuacji bardzo dużej liczby obserwacji w próbie, jest złudne na ogół przekonanie badacza o definitywnej redukcji niepewności wnioskowania dzięki bogatszej wiedzy empirycznej. Malejące rozproszenie statystyk próbkowych sugeruje, że całkowity błąd badania znacznie się zmniejsza. W rzeczywistości jednak tylko błąd losowy (ang. *sampling error*), będący funkcją liczebności próby, maleje proporcjonalnie do wzrostu liczby obserwacji. Błędy nielosowe (ang. *nonsampling errors*) prawie nigdy nie reagują tak na zwiększenie liczebności próby. A właśnie ta kategoria błędów, w której mieszczą się: błąd operatu losowania (błąd pokrycia), błąd braków odpowiedzi, różne błędy pomiaru i błąd przetwarzania danych, ma w praktyce badań społecznych i ekonomicznych coraz większe znaczenie. Nieświadomość konsekwencji tych błędów może zrodzić u badacza przekonanie nie tylko o dużej wiarygodności wnioskowania, lecz także o jego dużej precyzji. Wraz ze wzrostem liczebności próby zmniejsza się bowiem rozpiętość przedziałów ufności i rośnie moc testów statystycznych, czyli ich zdolność do rozróżnienia między hipotezą prawdziwą i fałszywą. Wystarczy jednak, że pewien bliżej nierozpoznany odsetek respondentów wylosowanych do próby odmówi udziału w badaniu i może się okazać, że prawdziwa wartość parametru nie mieści się w wąskich granicach przedziału ufności opartego na bardzo dużej próbie.

Podobne konsekwencje warto dostrzegać w przypadku testowania hipotez i operowania kategorią istotności statystycznej. W dużych próbach rozkład statystyki testowej, będącej syntetycznym miernikiem obserwacji zarejestrowanych w próbie, charakteryzuje się bardzo małą dyspersją, co w rezultacie prowadzi do częstego odrzucenia hipotezy zerowej[17]. Tylko dla wąskiego przedziału liczbowego możliwych wartości statystyki testowej indeks *p*-value przyjmuje wartości większe od minimalnych. Dowolnie małe odchylenie wartości statystyki w próbie od wartości oczekiwanej, jaką by otrzymano, gdyby hipoteza zerowa okazała się prawdziwa, może być – przy dużej liczebności próby – powodem uznania tego odchylenia za statystycznie istotne, a w konsekwencji do odrzucenia sprawdzanej hipotezy.

Dla badaczy różnych obszarów wiedzy stosujących metody wnioskowania statystycznego, w oczywisty sposób zainteresowanych odrzuceniem hipotezy zero-

[17] O potrzebie innego podejścia do testowania hipotez lub modyfikacji testów istotności w przypadku dużych liczebnie prób pisał już Kmenta (1990), proponując proste, aczkolwiek niedoskonałe rozwiązanie – zmianę poziomu istotności wraz z rosnącą wielkością próby, tak aby trudniej było odrzucić hipotezę zerową dla dużych prób.

wej, duża wielkość próby staje się bardzo pożądana. Pozyskanie znacznej liczby pomiarów w próbie jest współcześnie nieporównanie łatwiejsze niż w przeszłości. Niebezpieczeństwo tkwi w tym, że coraz częściej na dalszy plan schodzi w tych działaniach respektowanie założeń modelu wnioskowania statystycznego i rygorów próby losowej. Jeżeli zgodzilibyśmy się z tym, że „jesteśmy gotowi do poświęcenia odrobiny dokładności w zamian za poznanie ogólnego trendu" – jak deklarują Meyer-Schönberger i Cukier (2014, s. 44), autorzy głośnej książki *BIG DATA. Rewolucja, która zmieni nasze myślenie, pracę i życie* – byłoby to celowe i świadome zwiększanie ryzyka błędnych rozstrzygnięć przy użyciu metod wnioskowania statystycznego. Oznaczałoby dodanie do niepewności istniejącej w modelu wnioskowania jeszcze jednego elementu, który w zasadniczy sposób zakłócałby relacje opisane w założeniach modelu. Dlatego zwiększanie liczby obserwacji w próbie nie może zwalniać badacza z powinności dokładnego analizowania jakości danych.

Większa liczba informacji rzadko może zrekompensować ich niższą jakość. Dobry przykład stanowią odmowy respondentów. W wielu przypadkach powody odmów są współzależne z celami badania i z interesującymi badacza zmiennymi. A to – bez względu na wielkość próby – zawsze rodzi błędy systematyczne we wnioskowaniu. W testowaniu hipotez może to oznaczać przekonanie o bardzo niewielkim przedziale nierozstrzygnięcia (nieodrzucenia hipotezy zerowej), tyle że obszar ten będzie błędnie usytuowany na osi liczbowej. Wszystkie z wymienionych wcześniej błędów nielosowych, w tym m.in. błędy pokrycia i błędy pomiaru, mogą powodować obciążenia estymatorów, nieulegających zmniejszeniu pod wpływem nowych obserwacji. Zdaniem autora są to także te okoliczności, które sprowokowały świat nauki do dyskusji o istotności statystycznej i uproszczonej interpretacji dychotomicznej *p*-value. Nie można bowiem z jednej strony polegać na bardzo ostrej, jednoznacznej granicy liczbowej ($p < 0{,}05$), a z drugiej nie kontrolować w pełni stopnia spełnienia w konkretnym badaniu i dla konkretnej próby wszystkich założeń modelu statystycznego, za pomocą którego uzyskuje się wyniki pozwalające sformułować ostateczny wniosek z testowania hipotez. Stawką w dążeniach do uwzględnienia we wnioskowaniu wszystkich aspektów niepewności (nie tylko losowych) jest wiarygodność statystycznych badań niewyczerpujących i ich rola w dalszym rozwoju poznania naukowego.

Istnieje też druga, korzystniejsza strona zaangażowania technik big data w badaniach statystycznych. Rozumiane szeroko jako zasoby nowej wiedzy i sposoby jej zdobywania, wykorzystujące najnowsze sposoby gromadzenia i przetwarzania dużych zbiorów danych, mogą w znacznym stopniu przyczynić się do poprawy jakości badań próbkowych – estymacji i weryfikacji hipotez. W szczególności dotyczy to tych badań niewyczerpujących, w których znaczne są odsetki jednostek

populacji niepokrytych operatem, odmów udziału w badaniu lub błędnie udzielonych odpowiedzi (np. z winy projektanta badania bądź ankieterów)[18].

Pozapróbkowe źródła danych, aczkolwiek często trudne w analizie i przetwarzaniu ze względu na ich nieuporządkowanie lub nieustrukturyzowanie, tworzą już obecnie ważne dla statystyków metadane (ang. *metadata*) i paradane (ang. *paradata*). Metadane są informacjami, które opisują i wzbogacają dane statystyczne uzyskane w badaniu próbkowym, zapewniając ich właściwą interpretację. Zalicza się do nich informacje o wykorzystanych: instrumentach pomiarowych (np. kwestionariuszach), instrukcjach dla ankieterów, sposobach pomiaru sondażowego, programach do przetwarzania danych itp. Paradane zaś to dodatkowe informacje o gromadzeniu danych w próbie, takie jak obserwacje ankietera (np. dotyczące stopnia zainteresowania respondenta tematem badania), szczegółowe fakty (np. która kolejna próba kontaktu z respondentem okazała się skuteczna) oraz inne (np. czas, jakiego potrzebował respondent na udzielenie odpowiedzi na poszczególne pytania, a w ankiecie elektronicznej – czas między kliknięciami). Najbardziej jednak powszechnym obecnie sposobem walidacji i wzbogacania informacji próbkowych są znane od dawna techniki ważenia i kalibracji. Podstawę ich wykorzystania stanowią najczęściej inne, wcześniej zrealizowane badania statystyczne albo różnego rodzaju rejestry, w tym urzędowe.

Jakie jest miejsce wszystkich tych nowych rozwiązań epoki big data w testowaniu hipotez? Wydaje się, że mogą one stanowić fundament odpowiedzi na dylematy dyskutowane przez autorów wspomnianych dyskusji na forum „The American Statistician" oraz „Nature". Jeżeli uznajemy *p*-value za niewystarczający wskaźnik wagi przesłanek w próbie przeciw testowanej hipotezie, to głównie z dwóch powodów. Po pierwsze dlatego, że dość często istnieją wątpliwości co do jakości danych próbkowych i ich możliwego obciążenia błędami nielosowymi. Po drugie, trudno jest uznać jedną próbę losową za wystarczający materiał empiryczny do dedukcyjnego wnioskowania o nieprawdziwości hipotezy w warunkach, kiedy coraz częściej dostępne są inne (być może lepszej jakości niż próba) źródła danych o badanej populacji. Wykorzystanie danych spoza próby jawi się jako najbardziej obiecujące rozwiązanie wskazanych wątpliwości. Rodzaj takich danych, a także sposób ich wykorzystania są wówczas pierwszymi ważnymi wyzwaniami, przed jakimi staje statystyk.

W wielu sytuacjach pomocne w łączeniu wiedzy *a priori* i informacji z próby może się okazać twierdzenie Bayesa. Być może właśnie podejście bayesowskie lub paradygmat bayesowski, który opisuje proces aktualizacji wiedzy pod wpływem nowych informacji, stanie się przyszłością wnioskowania statystycznego?

[18] Poza badaniami niewyczerpującymi także w spisach (badaniach pełnych), które nie są wolne od błędów nielosowych, coraz szerzej korzysta się z dodatkowych źródeł informacji. W tym kontekście Gołata (2018) ogłasza i dobrze uzasadnia w swojej monografii koniec ery tradycyjnych spisów ludności.

Są tego pierwsze wyraźne oznaki w środowisku naukowym[19]. W naukach społecznych i ekonomicznych zaliczyć do nich należy zarówno rosnące przekonanie o konieczności posługiwania się personalistyczną (nieczęstościową) interpretacją prawdopodobieństwa we wnioskowaniu statystycznym, jak i zrozumienie dla dążenia do wykorzystania wielu źródeł informacji w podejmowaniu decyzji, także tych odnoszących się do odrzucenia lub nieodrzucenia testowanych hipotez.

PODSUMOWANIE

W dobie szybko rosnącej mocy obliczeniowej komputerów i łatwego w obsłudze oprogramowania statystycznego uwagę badaczy coraz częściej zajmuje przede wszystkim wynik analiz numerycznych. W przypadku badań niewyczerpujących jest nim – w szczególności w naukach społecznych i przyrodniczych – rezultat testowania określonych hipotez statystycznych. W rozstrzyganiu o tym, czy zaobserwowany efekt w próbie jest statystycznie istotny, rzadko informuje się czytelnika o ważnych uwarunkowaniach. Należy do nich zaliczyć informacje o: założeniach zastosowanego modelu statystycznego, jakości danych w próbie, liczbie i formie innych testowanych hipotez, a także o wykorzystanych narzędziach imputacji lub kalibracji brakujących danych. Bez tych informacji pojęcie istotności statystycznej jest mało wartościowe, bo trudno ocenić jego wiarygodność w odniesieniu do procesu wnioskowania.

Ocenę tej wiarygodności dodatkowo komplikuje nazbyt powszechny zwyczaj prezentowania rozstrzygnięć o odrzuceniu hipotezy zerowej wyłącznie na podstawie przekroczenia lub nieprzekroczenia przez wskaźnik *p*-value pewnego progu, zwykle określonego jako 0,05 (rzadziej jako 0,01). Takie dychotomiczne ujęcie ciągłej zmiennej losowej, jaką jest *p*-value, bez podania konkretnej wartości *p* w danej próbie, znacznie zubaża interpretację i możliwość oceny wyników testowania.

Nadmierne skupienie się badacza na wynikach komputerowych analiz statystycznych często skutkuje małą wnikliwością w sprawdzaniu założeń odnoszących się zarówno do stosowanego modelu wnioskowania, jak i do danych uzyskanych w próbie. I nie chodzi tu jedynie o sprawdzenie losowości próby, ale przede wszystkim o jakość danych z punktu widzenia ich obciążenia błędami nielosowymi. W tym kontekście big data, ze względu na swoje nieuporządkowanie i nieustrukturyzowanie, wymaga szczególnej czujności badacza. Równocześnie jednak zasoby big data dają szansę na wzbogacenie informacji próbkowych, zwłaszcza w sytuacji niskiego wskaźnika odpowiedzi respondentów lub słabej jakości narzędzi pomiarowych.

[19] Na przykład w odniesieniu do badań medycznych Ruberg i współpracownicy (2019, s. 320) stwierdzają: „The Bayesian way of thinking and formal analytical approach seems ideally suited for the drug development process" („Bayesowski sposób myślenia i formalne podejście analityczne wydają się idealnie dopasowane do prowadzenia badań nad rozwojem leków").

Łączenie informacji z różnych źródeł prawdopodobnie doprowadzi do wzrostu znaczenia indukcyjnego podejścia do wnioskowania statystycznego opartego na twierdzeniu Bayesa.

Załącznik – wyciąg z oświadczenia Amerykańskiego Towarzystwa Statystycznego na temat statystycznej istotności oraz *p*-value[20]

1. Wartości prawdopodobieństwa krytycznego (*p*-value) mogą wskazywać na to, jak nieprzystające do określonego modelu statystycznego są zaobserwowane dane.

P-value stanowi pewne podejście do syntetycznego wyrażenia niezgodności między określonym zbiorem danych a zaproponowanym modelem dla tych danych. Najczęstszym kontekstem, w jakim pojawia się *p*-value, jest model zbudowany przy określonych założeniach, łącznie z tzw. hipotezą zerową. Często hipoteza ta postuluje brak efektu, takiego jak różnica między dwiema zbiorowościami (albo średnimi dla tych zbiorowości), lub brak zależności pomiędzy analizowanym czynnikiem a uzyskanym wynikiem (np. zastosowaniem nowego leku a rezultatem wyrażonym pewnym miernikiem). Im mniejsze *p*-value, tym większa statystyczna niezgodność zaobserwowanych w próbie danych z hipotezą zerową, przy założeniu spełnienia założeń przyjętego na wstępie modelu.

2. *P*-value nie jest miarą prawdopodobieństwa tego, że analizowana hipoteza jest prawdziwa, ani tego, że dane zostały uzyskane wyłącznie w drodze losowania (zostały wygenerowane w procesie losowym).

Badacze często chcieliby przekształcić *p*-value w stwierdzenie dotyczące prawdziwości hipotezy zerowej albo w prawdopodobieństwo tego, że proces losowy wygenerował uzyskane dane. Trzeba podkreślić, że *p*-value nie jest ani jednym, ani drugim. Stanowi komunikat o danych w odniesieniu do pewnego określonego hipotetycznego wyjaśnienia, ale nie jest stwierdzeniem o tym wyjaśnieniu.

3. Konkluzje badawcze oraz decyzje ekonomiczne lub związane z określoną polityką działania nie powinny być oparte wyłącznie na tym, czy *p*-value przekroczyła określony próg.

Praktyka redukcji analizy danych lub naukowego wnioskowania do mechanicznej reguły „czerwonej linii" (takiej jak „$p < 0{,}05$") w celu uzasadnienia naukowych stwierdzeń lub wniosków może prowadzić do błędnych przekonań i złych

[20] Wyciąg ten obejmuje pkt 3, zatytułowany *Zasady* (*Principles*), który stanowi znaczną część oświadczenia. Za: Wasserstein i Lazar (2016, s. 131 i 132).

decyzji. Wniosek badawczy nie staje się natychmiast „prawdziwy" dlatego, że znalazł się po jednej stronie progu, lub „fałszywy", gdy ulokował się po drugiej stronie. Badacze powinni wziąć pod uwagę więcej czynników tworzących kontekst analizowanego problemu, takich jak: projekt całego badania, jakość dokonanych pomiarów, zewnętrzne źródła danych na temat badanego zagadnienia, a także spełnienie założeń, które tkwią u podstaw analizy danych. W praktyce decyzyjnej często wymagane jest binarne rozstrzygnięcie „tak – nie", ale nie oznacza to, że sama wartość *p* może przesądzić o tym, czy decyzja jest poprawna, czy nie. Szeroko stosowana kategoria „istotności statystycznej" (zwykle interpretowana jako „$p < 0{,}05$") jako licencji na uznanie wniosków naukowych (lub na sugerowanie prawdy) prowadzi do poważnego wypaczenia procesu badawczego.

4. Poprawne wnioskowanie wymaga od badacza ujawnienia pełnej informacji oraz przejrzystości.

Prezentowanie *p*-value i powiązanych z tym wskaźnikiem analiz nie może się odbywać wybiórczo. Prowadzenie złożonych, wielokrotnych analiz i przedstawianie jedynie tych z określonymi wartościami *p* (najczęściej przekraczającymi próg istotności) powoduje zasadniczo, że podane wielkości *p*-value są nieinterpretowalne. Przedstawianie jedynie obiecujących wyników, niczym wybieranie wisienek z tortu, znane też jako wybiórcze sięganie do danych, pogoń za istotnością, selektywne wnioskowanie lub nadużywanie założeń w celu osiągnięcia pożądanego *p*-value (*p-hacking*), prowadzi do sztucznej przewagi statystycznie istotnych rezultatów w publikowanych pracach naukowych. Do takich działań absolutnie nie powinno się dopuszczać. Nawet gdy formalnie nie prowadzi się wielokrotnego testowania, warto pamiętać, że jeżeli badacz – na podstawie uzyskanych wyników statystycznych – wybiera, co zaprezentować, to właściwa interpretacja wyników zostaje poważnie zagrożona, o ile czytelnikowi nie ujawnia się, że dokonano tego typu wyborów oraz nie informuje się o ich podstawach. Badacze powinni ujawniać liczbę rozważanych hipotez w realizowanym badaniu, wszelkie decyzje dotyczące uzyskiwania danych, wszystkie przeprowadzone analizy statystyczne, a także obliczone wskaźniki *p*-value. Nie można formułować wartościowych naukowo wniosków opartych na *p*-value i powiązanych statystykach bez informacji przynajmniej o tym, ile i jakie analizy zostały przeprowadzone oraz w jaki sposób dokonano wyboru niektórych z nich (wraz z *p*-value) do ostatecznego zaprezentowania.

5. *P*-value ani statystyczna istotność nie mierzą wielkości efektu ani ważności uzyskanego wyniku.

Statystyczna istotność nie jest równoważna z istotnością naukową, ludzką ani ekonomiczną. Mniejsze wartości *p* niekoniecznie oznaczają wystąpienie więk-

szych lub ważniejszych efektów [różnicy między tym, co zaobserwowano w próbie, a tym, czego należało oczekiwać, gdyby hipoteza zerowa była prawdziwa – przyp. autora]. Podobnie większe wartości *p* nie implikują nieistnienia ważnych efektów lub braku efektu w ogóle. Każdy efekt, niezależnie od tego, jak byłby niewielki, może generować małe *p*-value, jeżeli tylko liczebność próby lub precyzja pomiarów są wystarczająco duże. Analogicznie duży rozmiar efektów może generować duże *p*-value, gdy mała jest próba lub precyzja pomiarów. Podobnie zresztą identyczne oszacowania wielkości efektów będą miały różne wartości *p*, jeżeli różna będzie precyzja ich oszacowań.

6. Sam w sobie wskaźnik *p*-value nie stanowi dobrej miary przesłanek dotyczących modelu lub hipotezy.

Badacze powinni uznać, że wskaźnik *p*-value pozbawiony kontekstu lub innych przesłanek dostarcza jedynie ograniczonych informacji. Na przykład *p*-value bliskie 0,05 rozpatrywane w oderwaniu od kontekstu stanowi słaby dowód nieprawdziwości hipotezy zerowej. Podobnie zresztą jak stosunkowo duże *p*-value nie dowodzi prawdziwości hipotezy zerowej; wiele innych hipotez może być równie lub nawet bardziej zgodnych z danymi zaobserwowanymi w próbie. Z tych powodów analiza danych nie powinna kończyć się obliczeniem *p*-value w sytuacji, kiedy inne podejścia byłyby właściwe i możliwe do zastosowania.

BIBLIOGRAFIA

Amrhein, V., Greenland, S., McShane, B. (2019). Retire statistical significance. *Nature*, (567), 305–307.

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: sup 1), 262–270. DOI: 10.1080/00031305.2018.1543137.

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. (2018). Redefine Statistical Significance. *Nature Human Behaviour*, (1), 6–10.

Beręsewicz, M., Szymkowiak, M. (2015). Big data w statystyce publicznej – nadzieje, osiągnięcia, wyzwania i zagrożenia. *Ekonometria*, (2), 9–22.

de Finetti, B. (1964). Foresight: Its Logical Laws, its Subjective Sources. W: H. E. Kyburg, Jr., H. E. Smokler (red.), *Studies in Subjective Probability* (s. 93–158). New York: Wiley.

Fisher, R. A. (1925). *Statistical Methods for Research Workers*. Edinburgh: Oliver & Boyd.

Gelman, A., Stern, H. (2006). The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician*, (4), 328–331.

Gołata, E. (2018). *Koniec ery tradycyjnych spisów ludności*. Poznań: Wydawnictwo Uniwersytetu Ekonomicznego w Poznaniu.

Goodman, S. N. (1999). Toward evidence-based medical statistics. 1: The *P* value fallacy. *Annals of Internal Medicine*, (12), 1005–1013.

Goodman, S. N. (2019). Why is Getting Rid of *P*-Values So Hard? Musings on Science and Statistics. *The American Statistician*, (73: sup 1), 352–357.

Hurlbert, S. H., Levine, R. A., Utts, J. (2019). Coup de Grâce for a Tough Old Bull: "Statistically Significant" Expires. *The American Statistician*, (73: sup 1), 26–30.

Kmenta, J. (1990). *Elements of econometrics*. New York: Macmillan Publishing Company.

Mayer-Schönberger, V., Cukier, K. (2014). *Big data. Rewolucja, która zmieni nasze myślenie, pracę i życie*. Warszawa: MT Biznes.

Ramsey, F. P. (1964). Truth and Probability. W: H. E. Kyburg, Jr., H. E. Smokler (red.), *Studies in Subjective Probability* (s. 63–92). New York: Wiley.

Ruberg, S. J., Harrell, F. E. Jr., Gamalo-Siebers, M., LaVange, L., Lee, J. J., Price, K., Peck, C. (2019). Inference and Decision Making for 21st-Century Drug Development and Approval. *The American Statistician*, (73), 319–327.

Savage, L. J. (1954). *The Foundations of Statistics*. New York: Wiley.

Stefanowicz, B., Cierpiał-Wolan, M. (2015). Błędy przetwarzania danych. *Wiadomości Statystyczne*, (9), 23–29.

Szreder, M. (2004). Od klasycznej do częstościowej i personalistycznej interpretacji prawdopodobieństwa. *Wiadomości Statystyczne*, (8), 1–10.

Szreder, M. (2010). O weryfikacji i falsyfikacji hipotez. *Przegląd Statystyczny*, (2–3), 82–88.

Szreder, M. (2013). Twierdzenie Bayesa po 250 latach. *Wiadomości Statystyczne*, (12), 23–26.

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

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.

Wasserstein, R. L., Schirm, A. L., Lazar, N. A. (2019). Moving to a World Beyond "p < 0.05". *The American Statistician*, (73: sup 1), 1–19.