Mirosław Szreder

Rola badań statystycznych w naukach ekonomicznych w świetle nowych możliwości określanych mianem *big data*

Wprowadzenie

Rozwój technologiczny, którego doświadczamy od kilkunastu lat, tworzy nowe, nieznane wcześniej możliwości przetwarzania dużych ilości informacji (nie tylko liczbowych) i wzbogacania na tej podstawie wiedzy o otaczającym nas świecie. Poszukiwanie tego rodzaju wiedzy i wydobywanie jej ze zbiorów danych liczbowych było od dziesiątków lat głównym przedmiotem zainteresowania statystyków. Obecnie, gdy na wszystkich etapach badania statystycznego coraz większą rolę odgrywają technologie informatyczne i wyrafinowane narzędzia sztucznej inteligencji (w tym silnie rozpowszechnione w ostatniej dekadzie sztuczne sieci neuronowe), zmianie musi ulegać też sama koncepcja badań statystycznych, a także klasyczna – znana z programów uniwersyteckich – metodyka statystyki jako nauki. Samą profesję statystyka zastępuje coraz częściej analityk danych, co ma wskazywać na wykraczanie współczesnych analiz danych ilościowych i jakościowych poza sferę tradycyjnie rozumianej statystyki.

We współczesnym podejściu do badań statystycznych pojawiają się pewne nowe tendencje, z których większość koncentruje się na szybszym i wszechstronniejszym poznawaniu analizowanego fragmentu rzeczywistości. Nie wszystkie one muszą się jednak okazać korzystne i trwałe. W niniejszym opracowaniu podjęto próbę scharakteryzowania najważniejszych zmian w podejściu do badań statystycznych i ich zastosowań w ekonomii. W charakterystyce tej uwzględnione zostaną zarówno nowe możliwości, jakie zmiany te oferują, jak również wyzwania i zagrożenia, które podjąć będą musieli statystycy i ekonomiści w najbliższej przyszłości.

1. Wzrost znaczenia badań reprezentacyjnych

W empirycznych zastosowaniach statystyki jedną z najbardziej trwałych tendencji, kształtujących się już od kilku dekad jest zastępowanie badań pełnych (wyczerpujących) badaniami próbkowymi, w tym przede wszystkim badaniami reprezentacyjnymi – opartymi na próbach losowych. Spowodowane jest to nie tylko postępami technologii komputerowej, umożliwiającej szybkie zbieranie, przetwarzanie i analizowanie danych. Równie ważnymi przesłankami tych zmian są postępy w rozwoju teorii próbkowania i

wnioskowania statystycznego, a także rosnąca powszechnie świadomość niedostatków badań wyczerpujących. Ten ostatni czynnik – błędy i słabości badań wyczerpujących (spisów) – był w przeszłości rzadko dyskutowany. W wielu środowiskach (np. dziennikarskim) nadal pokutuje przekonanie, iż wiarygodny i dokładny opis cech danej zbiorowości zapewnia jedynie badanie pełne, czyli objęcie pomiarem wszystkich jednostek populacji. Taki wyidealizowany obraz badania pełnego jest zwykle skutkiem zbyt wąskiego rozumienia natury i złożoności błędu, jakim mogą być obciążone wyniki badania statystycznego. Utożsamianie całkowitego błędu w badaniach niewyczerpujących z błędem losowania, co powszechnie się sugeruje w większości komunikatów z badań sondażowych, prowadzi do błędnego przekonania, że w badaniach wyczerpujących (pełnych) błędy w ogóle nie występują. Tymczasem praktyka obu rodzajów badań – pełnych i próbkowych – wskazuje dość jednoznacznie, że w strukturze całkowitego błędu badania coraz większy (a często dominujący) jest udział tych składników, które z losowaniem próby mają niewiele wspólnego[1]. Zaliczyć do nich należy przede wszystkim błędy spowodowane:

- niekompletnym lub złej jakości operatem losowania (błąd pokrycia, ang. *coverage error*)[2],
- odmowami respondentów udziału w badaniu (ang. *nonresponse error*)[3],
- zarejestrowaniem nieprawdziwych danych z winy ankietera lub respondenta (błąd pomiaru, ang. *measurement error*),
- złym przetwarzaniem zgromadzonych danych (ang. *postsurvey processing error*)[4].

Wszystkie wyliczone wyżej rodzaje błędów składają się na kategorię błędu nielosowego, zniekształcającego zarówno wyniki badań wyczerpujących, jak i niewyczerpujących. Świadomość występowania tych błędów w badaniach pełnych, w tym także w ogólnokrajowych spisach powszechnych, prowadziła – wraz z postępami w zakresie

[1] Szerzej na ten temat por. [Szreder, 2015].

[2] Błąd ten występuje we wszystkich rodzajach badań, także tych realizowanych przez statystykę publiczną. Potwierdza to m.in. [Paradysz, 2010, s. 52], który pisze: „Wydaje się, że obecność błędów pokrycia w badaniach statystycznych GUS była dostrzegana dość dawno, ale stanowiły one problem wstydliwy, o którym oficjalnie nie pisano". Z kolei o pominięciu 900 tys. mężczyzn w wieku poniżej 40 lat w narodowym spisie powszechnym w W. Brytanii w roku 2001 pisze [Henley, 2011].

[3] [Särndal i Lundström, 2006] stwierdzają: „współcześnie braki odpowiedzi są normalną (chociaż niepożądaną) cechą badań ankietowych". A [Donsbach i Traugott, 2008, s. 304] dodają: "Z pewnymi wyjątkami wskaźniki odpowiedzi (*response rates*) rzadko przewyższają obecnie 50%. Nawet dobrym organizacjom badań społecznych trudno jest dzisiaj osiągnąć 50% wskaźniki odpowiedz".

[4] Obszernie o specyfice tego błędu piszą m.in. [Stefanowicz i Cierpiał-Wolan, 2015].

próbkowania i wnioskowania (czym zajmuje się metoda reprezentacyjna) – do ukształtowania się trwałej, nieprzerwanej aż do czasu pojawienia się koncepcji *big data* tendencji do zastępowania badań pełnych badaniami reprezentacyjnymi.

Do znanych wcześniej zalet badań próbkowych – niskich relatywnie kosztów realizacji, krótkiego czasu ich wykonywania oraz przetwarzania danych – zaliczyć trzeba łatwiejsze niż w przeszłości sposoby elektronicznego kontaktu z respondentami, sprawnego komputerowego zbierania danych, ich przetwarzania i analizy. W wielu zagadnieniach ekonomicznych i społecznych badania reprezentacyjne stały się na tyle popularne – głównie z racji wspomnianych ułatwień w ich projektowaniu i realizacji – iż coraz częściej wypierają one bardziej czasochłonne, lecz właściwsze w konkretnej problematyce badania jakościowe. Pokusa użycia prostego kwestionariusza ze wszystkimi pytaniami zamkniętymi, jako jedynego narzędzia badawczego, jest dla wielu naukowców, zwłaszcza młodego pokolenia, trudna do przezwyciężenia. Można by podać wiele przykładów publikacji, w których motywy postępowania osób badanych (w rzeczywistości często nieuświadomione) lub ich postawy w określonych okolicznościach, analizuje się wyłącznie na podstawie udzielonych przez nich odpowiedzi w badaniu ankietowym. Tam, gdzie powinny być użyte takie techniki, jak: indywidualne wywiady pogłębione, zogniskowane wywiady grupowe, czy też techniki projekcyjne, wykorzystany zostaje wyłącznie prosty kwestionariusz. Jeśli doda się do tego częste błędy badacza, odnoszące się do wyboru operatu losowania, techniki próbkowania, czy też sposobu pomiaru sondażowego, to dziwić nie powinna pogarszająca się reputacja statystycznych badań próbkowych, niezależnie od dziedziny ich zastosowań. W finansach na przykład, gdzie badań ankietowych pojawia się coraz więcej, do uzyskanych z nich wyników, a także prawidłowości, na które zdają się one wskazywać, część czasopism podchodzi z rezerwą. Autorzy opracowania [Baker i Mukherjee, 2007] zapytali o rolę badań ankietowych w publikacjach z zakresu finansów pięćdziesięciu redaktorów najpopularniejszych w świecie czasopism naukowych z tego obszaru. Wyniki swojego badania[5] ujęli w podziale na odpowiedzi udzielone przez redaktorów reprezentujących kluczowe czasopisma (ang. *core journals*)[6] i pozostałe (ang. *non-core journals*)[7], por. Tablica 1.

[5] Samo badanie także było ankietowe, a jego autorzy doświadczyli typowego dla tego typu badań niskiego wskaźnika odpowiedzi. Udzieliło ich jedynie 23 spośród 50 redaktorów czasopism.
[6] Znalazły się wśród nich m.in.: „Financial Review", "Journal of Finance", "Journal of Financial Economics", "Review of Financial Studies".
[7] Są wśród nich m.in.: "Applied Financial Economics", "European Journal of Finance", "International Review of Economics and Finance", "Quarterly Review of Economics and Finance".

Tablica 1. Rola badań ankietowych w czasopismach naukowych z zakresu finansów – opinie redaktorów czasopism

Które z następujących stwierdzeń najlepiej opisuje rolę, jaką powinny odgrywać badania próbkowe w literaturze z zakresu finansów	Redaktorzy kluczowych czasopism	Redaktorzy pozostałych czasopism	Ogółem (n)	Ogółem (%)
A. Badanie próbkowe powinno być traktowane na równi z innymi oryginalnymi badaniami.	-	10	10	43,5
B. Badanie próbkowe powinno pełnić rolę uzupełniającą względem innego oryginalnego badania.	4	6	10	43,5
C. Rola badania próbkowego jest ograniczona (lub nie ma ono żadnego znaczenia) w stosunku do innych oryginalnych badań.	2	1	3	13,0
D. Rola badania próbkowego powinna być następująca (podaj): …	-	-	-	-

Źrodło: [Baker i Mukherjee, 2007, s. 21].

Redakcje kluczowych czasopism z zakresu finansów nie są skłonne traktować opracowań opartych wyłącznie na badaniach próbkowych jako równorzędnych z innymi opracowaniami naukowymi. Przypisują im rolę uzupełniającą, albo wskazują na ich ograniczone znaczenie. W mniej prestiżowych czasopismach badania te traktuje się w większości albo jako równorzędne z innymi rodzajami badań, albo jako uzupełnienie innej metodyki badawczej. Trudno jest jednoznacznie stwierdzić, czy traktowanie rozważań naukowych opartych badaniach ankietowych jako jedynie uzupełniających w stosunku do głównego nurtu badań, wiąże się z ich naturalnymi ograniczeniami (o których wspomniano wcześniej), czy też jest skutkiem pogarszającej się reputacji badań ankietowych. Ich masowość bowiem, z którą wiąże się często nieprzygotowanie lub brak kompetencji ich realizatorów, powoduje nie tylko wrażenie nadmiaru ankiet i sondaży w wielu dziedzinach nauki, ale oznacza także niższą niż w przeszłości ich jakość. Mamy więc z jednej strony bardzo wysokiej jakości badania reprezentacyjne, skrupulatnie zweryfikowane na podstawie pomiarów wszystkich jednostek populacji (np. *exit poll* podczas wyborów[8]), a z

[8] Od roku 2010 we wszystkich ogólnokrajowych wyborach powszechnych w Polsce notuje się dużą dokładność w badaniach *exit poll* – badaniach reprezentacyjnych, w których pomiarem objętych zostaje mniej niż 1% populacji (biorących udział w glosowaniu wyborców). Szczegóły por. m.in. [Szreder, 2016].

drugiej – dużą liczbę badań ankietowych słabej jakości, podważających wartość samej metody badawczej, a nawet więcej – statystyki jako nauki.

Z tego między innymi powodu, badania próbkowe nie stanowią wyłącznej alternatywy dla mało już popularnych badań wyczerpujących w empirycznej statystyce. Inną możliwość stanowią dane pochodzące z rejestrów urzędowych, zdolne do zastąpienia części danych spisowych o gospodarstwach domowych, osobach indywidualnych, czy przedsiębiorstwach. Narodowe spisy powszechne, ze swoją ponad 200-letnią historią, w coraz większym stopniu korzystają z wielu danych administracyjnych, a w konsekwencji są narażone na podnoszone coraz częściej wątpliwości, co do celowości ich dalszego trwania. Jest to jedynie jeden z przykładów innej, ogólniejszej tendencji obserwowanej w badaniach statystycznych, mianowicie coraz częstszego korzystania z kombinacji wielu źródeł informacji.

2. Integracja źródeł danych statystycznych

W praktyce badań niewyczerpujących od dawna starano się korzystać z więcej niż jednego źródła informacji. Wymagała tego albo teoria wnioskowania statystycznego, w szczególności paradygmat Bayesowski[9], albo konkretna technika próbkowania, której jednym z założeń jest posiadanie wiedzy wstępnej o populacji[10]. Nigdy wcześniej jednak działania zmierzające do integracji, czyli łącznego wykorzystania wielu źródeł danych w badaniach statystycznych, nie były tak powszechne jak obecnie. Przyczyny tego tkwią zarówno po stronie popytu – zapotrzebowania na informacje spoza próby, jaki po stronie podaży – łatwiejszej niż obecnie dostępności do tzw. „nowych źródeł danych"[11].

Jednym z najważniejszych czynników odpowiedzialnych za rosnący popyt na dodatkowe informacje, a w konsekwencji na integrację wielu źródeł danych, jest dążenie statystyków do zmniejszenie skutków rosnącej skali i znaczenia błędów nielosowych. Wymienione w poprzednim punkcie najważniejsze elementy tego błędu, począwszy od błędu pokrycia, a skończywszy na błędzie przetwarzania danych, stały się w wielu badaniach reprezentacyjnych znacznie bardziej kłopotliwe od błędu losowania. Temu

[9] Zakłada się w nim istnienie dwóch źródeł informacji: tak zwanej wiedzy *a priori* badacza oraz informacji z próby losowej. Wnioskowanie opiera się na połączeniu tych dwu źródeł za pomocą znanego twierdzenia Bayesa, por. np. [Szreder, 2013].

[10] Jednym z takich schematów jest popularne w praktyce losowanie warstwowe, w którym wiedzę spoza próby wykorzystuje się do powarstwowania populacji na bardziej jednorodne subpopulacje w stosunku do całej badanej zbiorowości.

[11] Terminu tego używa GUS, m.in. w opracowaniu „Statystyka publiczna – współczesne oblicze", [GUS, 2015].

ostatniemu – zdaniem wybitnego statystyka Leslie'go Kisha poświęcano w przeszłości zbyt wiele uwagi, zaniedbując prace nad błędami o charakterze nielosowym. Stwierdzenie "*sampling error is «over-researched«*"[12] (błąd losowania jest nadmiernie badany) było już przed laty apelem L. Kisha skierowanym do statystyków o większy wysiłek badawczy nad nielosowymi składnikami całkowitego błędu badania próbkowego. Składniki te jednak trudno poddają się modelowaniu czy opisowi probabilistycznemu, stąd wolniejsze były przez lata postępy w teoretycznym podejściu do nich, aniżeli w praktycznych działaniach ograniczających ich negatywne oddziaływanie, np. błędów braków odpowiedzi. W ostatnim okresie jednak wypracowanych zostało wiele metod i technik służących uzupełnieniu lub wzbogaceniu niedoskonałej wiedzy z próby, w większości opartych na założeniu dostępności badacza do informacji spoza próby (*a priori*) lub do próbkowej informacji o innych cechach badanych jednostek. Integracja kilku źródeł danych stała się obecnie niemal obowiązującą praktyką.

Praktyczne aspekty wykorzystania w badaniu statystycznym kombinacji kilku źródeł danych są zwykle analizowane indywidualnie w każdym konkretnym badaniu. Wspólną bowiem cechą integracji danych pochodzących z różnych źródeł jest brak uznanego powszechnie podejścia teoretycznego do wnioskowania na ich podstawie. Odnosi się to zarówno do danych administracyjnych (ang. *register-based statistics*), jak i do innych źródeł, związanych przede wszystkim z rozwojem nowych technologii: danych z portali społecznościowych, informacji określanych terminami *metadata*, *paradata* i innych. Z jednej strony wszystkie one stanowią znaczny potencjał informacyjny, mogący uzupełnić dostępne dane (np. z próby) lub zastąpić brakujące obserwacje, a z drugiej pozostają wciąż wyzwaniem dla statystyków, jeśli chodzi o ramy metodyczne ich integracji i dalszego wykorzystania.

Co prawda przez ostanie 10 lat dokonał się pewien postęp w zakresie korzystania przez statystykę (zwłaszcza tę oficjalną) z danych administracyjnych, ale w znaczne j mierze aktualne pozostaje spostrzeżenie autorów książki pt. *Register-based Statistics. Administrative Data for Statistical Purposes* – [Wallgren A. i Wallgren B., 2007], iż brak jest ugruntowanego i powszechnie akceptowanego podejścia teoretycznego w tym obszarze[13]. W sferze praktycznej postęp dokonuje się szybciej, przede wszystkim w

[12] Sformułowanie to pojawiło się m.in. w artykule Richarda Platka i Carla-Erika Särndala pt. "*Can a statistician deliver?*" opublikowanym w jęz. polskim wraz z dyskusją przez czasopismo naukowe "Wiadomości Statystyczne", por. [Platek i Särndal, 2001].

[13] „*Although register-based statistics are the most common form of statistics, no well-established theory in the field has existed up to now*" [Wallgren A. i Wallgren B., 2007, s. IX].

doskonaleniu systemów weryfikacji jakości danych w rejestrach urzędowych, a także w organizacyjnej i informatycznej synchronizacją działań urzędów administracji i organów statystyki publicznej. Główny Urząd Statystyczny twierdzi, że już w 66% badań statystycznych wykorzystuje tzw. nowe źródła danych, w tym przede wszystkim rejestry urzędowe. Do najważniejszych źródeł danych administracyjnych, z których GUS korzysta należą systemy: ewidencji ludności, podatkowy, informacji o działalności gospodarczej, o działalności rolniczej, o środowisku, zabezpieczenia społecznego, ubezpieczeń społecznych, ubezpieczeń zdrowotnych, informacji o przestępczości i wymiarze sprawiedliwości, informacji o edukacji, informacji o nieruchomościach, informacji o pojazdach i ich właścicielach, por. [GUS, 2015, s. 34]. Dla przyszłości badań ekonomicznych tendencja ta – otwartości statystyki publicznej na zbiory danych rejestrowane i gromadzone przez inne podmioty – jest obiecującą perspektywą. Może się ona przyczynić zarówno do poprawy jakości samych wyników badań statystycznych, jak i do zwiększenia dostępności do aktualnych (mniej opóźnionych) danych statystyki publicznej.

Terminem *metadata* (w jęz. polskim metadane) określa się ogół informacji o zgromadzonych danych statystycznych (ang. „*data about the data*"), czyli w szczególności ich strukturę, zakres i kontekst. Do metadanych zaliczyć należy informacje o wykorzystanych: instrumentach pomiarowych (np. kwestionariuszach), instrukcjach dla ankieterów, sposobach pomiaru sondażowego, programach do przetwarzania danych, itp. Informacje te ułatwiają ocenę wiarygodności i poprawności zebranych danych, a ponadto dostarczają szerszego kontekstu do interpretacji zgromadzonych lub zgromadzonych i przetworzonych danych[14].

Część badaczy, niezależnie od szerokiego zakresu metadanych, wyodrębnia dodatkowo zbiór szczegółowych, na ogół trudnych do zarejestrowania, lecz użytecznych informacji o badaniu, nazywając je *paradata*[15]. Są to informacje, które określić można jako towarzyszące samemu badaniu, mogące przyczynić się do lepszej ewaluacji jakości pozyskanych danych. Zalicza się do nich obserwacje ankietera (np. dotyczące stopnia zainteresowania respondenta tematem badania), szczegółowe fakty (np. która kolejna próba kontaktu z respondentem okazała się skuteczna), oraz inne dane (np. czas, jaki potrzebował respondent na odpowiedzi w poszczególnych pytaniach – czas między kliknięciami w ankiecie komputerowej). Są to informacje zarówno o charakterze

[14] Szerzej na temat roli metadanych w statystyce patrz: [Dippo i Sundgren, 2000].
[15] Szerzej o specyfice paradata por. [Kreuter, 2015].

obiektywnym, jak i subiektywnym. Subiektywne obserwacje dokonane przez ankietera mogą służyć do skonstruowania subiektywnych wag probabilistycznych, odzwierciedlających wiarygodność podanych przez respondenta informacji. Podejście takie, nieco bardziej sformalizowane, stosowane jest od wielu lat w przedwyborczych badaniach sondażowych i nazywane jest „*likely voter technique*" (ocena prawdopodobieństwa wzięcia udziału w wyborach przez respondenta wylosowanego z populacji wszystkich uprawnionych do głosowania)[16]. Obie omówione wyżej kategorie danych – *metadata* i *paradat*a – czekają jednak na wypracowanie przez statystyków spójnego podejścia metodycznego do ich zastosowań w badaniach statystycznych.

3. *Big data* a tradycyjne badania statystyczne

Big data jest znacznie szerszą kategorią od scharakteryzowanych w poprzednim punkcie urzędowych i nieurzędowych źródeł danych. Określa się nią taki sposób zdobywania nowej wiedzy i poznawania rzeczywistości, który może być zrealizowany w dużej skali, dzięki nowym możliwościom gromadzenia i przetwarzania wielkich zbiorów danych. Analityczna strona *big data* sprowadza się przede wszystkim do badania powiązań, współzależności i korelacji. Najszersze praktyczne zastosowanie znajduje zaś w przewidywaniu (prognozowaniu) różnych zjawisk i zachowań w niemal wszystkich dziedzinach życia. W przewidywaniach tych istotną zaletą *big data* jest możliwość bieżącej aktualizacji predyktywnych statystyk w oparciu o napływające nowe dane. Ogół operacji, także tych związanych z samouczeniem się algorytmów, odbywa się bez opóźnień – w czasie rzeczywistym. Czy w tej nowej rzeczywistości jest jeszcze miejsce na tradycyjne badania statystyczne, w szczególności badania reprezentacyjne?

W przeciwieństwie do autorów głośnej książki [Mayer-Schönberger i Cukier 2014], którzy twierdzą, że „*idea badania próbek traci sens, skoro możemy korzystać z dużej liczby danych*" (s. 50), uważam, że jeszcze przez długi okres czasu będą w badaniach ekonomicznych obecne zarówno badania próbkowe, jaki analizy i prognozy *big data*. Wydaje się bowiem, że najbardziej prawdopodobny jest równoległy rozwój dwóch tendencji – korzystania przez statystyków z możliwości big data, w celu wzbogacenia informacji próbkowej w badaniach reprezentacyjnych, oraz stopniowego zastępowania w niektórych dziedzinach tradycyjnych badań statystycznych wynikami algorytmów *big data*.

[16] Szerzej na ten temat pisze m.in. [Szreder 2007, 2010].

Obie te tendencje są współcześnie widoczne w badaniach ekonomicznych i społecznych. Pierwsza z nich, która rolę *big data* widzi jako komplementarną w stosunku do badań próbkowych, jest odpowiedzią na coraz większy udział nielosowych komponentów w całkowitym błędzie badania próbkowego. *Big data* może stanowić źródło wartościowej wiedzy potrzebnej do: imputacji brakujących danych, weryfikacji operatu losowania, korygowania struktury próby przy użyciu technik imputacji i kalibracji. Technologie *big data* mogą być także użyte do zgromadzenia i przetworzenia danych mogących poprawić jakość wnioskowania, np. scharakteryzowanych wcześniej zbiorów *metadata* i *paradata*.

Druga tendencja – zastępowanie przez big data badań próbkowych nie będzie miała charakteru rewolucyjnego, jak w tytule swojej monografii wieszczą [Mayer-Schönberger i Cukier 2014], z kilku powodów. Po pierwsze, w wielu dziedzinach ekonomii, a także życia społecznego, ważne będzie poznanie nie tylko ogólnego obrazu, albo współzależności cech, ale precyzyjne określenie charakterystyk populacji. Innymi słowy, ekonomiści nie zawsze zgodzą się z sugestią wspomnianych autorów, że: „*jesteśmy gotowi do poświęcenia odrobiny dokładności w zamian za poznanie ogólnego trendu*" (s. 55). Po drugie, nie zawsze badacze zadowolą się poznaniem związków korelacyjnych, bardzo użytecznych do prognozowania, ale mniej wartościowych w wyjaśnianiu zjawisk. Szereg korelacji między zmiennymi może nie mieć żadnego logicznego uzasadnienia (ang. *spurious correlations*) lub okazać się mało przydatnymi w wyjaśnianiu przyczyn zjawisk. „*W big data ważna jest odpowiedź na pytanie, co się dzieje, a nie dlaczego. Nie zawsze musimy znać przyczyny jakiegoś zjawiska, możemy po prostu pozwolić danym mówić za siebie*" – piszą [Mayer-Schönberger i Cukier 2014, s. 30]. Ale w naturze człowieka, nie tylko naukowca, leży dążenie do zrozumienia, do poznania przyczyny, a więc trudno sobie wyobrazić, że *big data* będzie w stanie wyeliminować inne rodzaje badań. Po trzecie, warto pamiętać, że duża ilość informacji, nie zawsze przekłada się na poprawę jakości opisu lub wnioskowania statystycznego. Dobrze zaprojektowane i zrealizowane badanie reprezentacyjne może precyzyjniej opisywać dany fragment rzeczywistości, aniżeli duża ilość informacji zebranych w sposób przypadkowy lub mało uporządkowany. Pewien zaś nieład jest charakterystyczny właśnie dla *big data*. I często w tych warunkach niewystarczający dla poprawnego opisu zbiorowości może się okazać zbiór obserwacji dokonanych na 90% lub nawet 99% jednostek.

Innym istotnym powodem tego, że badania próbkowe będą prawdopodobnie współistnieć równolegle z rozwojem sztucznej inteligencji i innych sposobów eksploracji

dużych zbiorów danych liczbowych jest brak adekwatnego podejścia metodologicznego do wnioskowania statystycznego w sytuacji posiadania próby losowej o bardzo dużych rozmiarach. Wyzwanie to można stosunkowo łatwo dostrzec w klasycznej teorii weryfikacji hipotez. Podkreśla się w niej m.in. pozytywny wpływ rosnącej liczebności próby na moc testu, czyli na zdolność do prawidłowego rozstrzygania o prawdziwości lub nieprawdziwości testowanej hipotezy. Jednak – co warto zauważyć – oba błędy, jakie w testowaniu hipotez bierze się pod uwagę (pierwszego i drugiego rodzaju), są związane z losowością próby, a dokładniej z niedoskonałością mechanizmu generującego obserwacje losowe z danego rozkładu (z danej populacji). Nie uwzględnia się innych błędów, o charakterze nielosowym. Z jednej strony więc maleje wraz ze wzrostem liczebności próby dyspersja rozkładu statystyki testowej, co w warunkach braku zakłóceń spowodowanych czynnikami nielosowymi oznacza lepszą moc testu. Z drugiej zaś, przy bardzo małej dyspersji rozkładu statystyki z próby (a w konsekwencji małym obszarze nieodrzucenia hipotezy), każde nawet niewielkie uchybienie pomiarowe, albo inny błąd systematyczny, prowadzić będą do odrzucenia hipotezy zerowej, gdyż wartość statystyki łatwo znajdzie się w rozległym obszarze krytycznym. Duże liczebnie próby powodują więc większą wrażliwość testu na błędy o charakterze nielosowym, w tym w szczególności na błędy systematyczne. Problem ten jest jednak głębszy i nie dotyczy wyłącznie błędów nielosowych. Gdyby abstrahować od tej kategorii błędów, a skoncentrować się tylko na błędzie losowania, to i wówczas duże liczebności prób stanowią wyzwanie dla podjęcia prawidłowej decyzji w oparciu o test statystyczny. W szczególności trudne do utrzymania są w tych okolicznościach typowe poziomy istotności 1% lub 5%. Przy bardzo małym rozproszeniu statystyki testowej, towarzyszącej dużej liczebności próby, typowe poziomy istotności będą determinowały na tyle duże obszary krytyczne, że prawie zawsze podjęta zostanie decyzja o odrzuceniu hipotezy zerowej. Świadomość konieczności wzięcia pod uwagę wielkości próby przy ustalaniu poziomu istotności istniała wśród statystyków od dawna. Zwracał na nią uwagę m.in. [Kmenta, 1990], proponując najprostsze, chociaż niedoskonałe rozwiązanie: zmianę poziomu istotności wraz z rosnącą wielkością próby, tak aby trudniej było odrzucić hipotezę zerową dla dużych prób niż dla małych. Współcześnie badacze rozpatrują ten sam problem, tyle że raczej w kategoriach *p-value*, aniżeli poziomu istotności, por. np. [Goldfarb i Lu, 2006], [Greene, 2003], [Hubbard i Armstrong, 2006], [Lin, Lucas i Shmueli, 2013]. Autorzy ostatniej z wymienionych pozycji uzasadniają, że z wyjątkiem sytuacji, kiedy parametr populacji jest dokładnie równy wartości ujętej w hipotezie zerowej (co w praktyce zdarza się bardzo rzadko), przy

rosnącej do nieskończoności wielkości próby *p-value* dążyć będzie do zera (co oznaczać będzie, że prawie zawsze odrzucona zostanie hipoteza zerowa). Wynika to wprost z asymptotycznej własności estymatorów wykorzystywanych w statystyce testowej, jaką jest zgodność. W konsekwencji dla bardzo dużych prób, klasyczne testy istotności nie są już skutecznym narzędziem rozstrzygającym o prawdziwości lub nieprawdziwości hipotezy statystycznej. Potrzebne jest nowe podejście. Powinno ono uwzględniać nie tylko duże rozmiary próby, ale także – co często charakteryzuje *big data* – odstępstwa od założenia o losowości próby. Dysponując ogromnymi zbiorami obserwacji oczekuje się, że duża ilość danych próbkowych zdoła zrekompensować brak losowości w mechanizmie odpowiedzialnym za ich generowanie. Dlatego obecnie wydobywanie wiedzy z wielkich zbiorów danych silniej wiąże się z technikami sztucznej inteligencji, wspieranymi przez nowe rozwiązania technologiczne, aniżeli z klasycznym wnioskowaniem statystycznym. To ostanie odgrywa nadal istotną rolę w badaniach reprezentacyjnych, a także w badaniach wykorzystujących zintegrowane źródła danych.

Podsumowanie

Tempo, w jakim powstają nowe sposoby pozyskiwania, gromadzenia i przetwarzania danych liczbowych sprawia, że z jednej strony statystyka występuje w coraz większej liczbie rożnych badań ekonomicznych, a z drugiej jej metody i techniki stają się niewystarczające wobec nowych wyzwań, w tym przede wszystkim możliwości oferowanych przez *big data*. Rosnącą popularność badań próbkowych w naukach ekonomicznych i społecznych odczytywać można jako potwierdzenie użyteczności wnioskowania statystycznego w poznawaniu wielu różnych zagadnień. Jednak wtedy, gdy badania te zaczynają wypierać właściwsze w danym problemie badawczym metody jakościowe, statystyka bywa niesprawiedliwie krytykowana za coś, co stanowi jej naturalne ograniczenie. Pewien przerost udziału badań statystycznych w poznawaniu rzeczywistości, i towarzyszące jemu nadmierne uproszczenie pomiarów powodują, że jakość części badań próbkowych pogarsza się. Tak jest m.in. z badaniami opinii publicznej i z niektórymi badaniami marketingowymi. Dodatkowo, niektórzy badacze nie są w pełni świadomi rosnącej wagi błędów nielosowych, jakimi obciążone bywają uzyskane przez nich wyniki.

Profesjonalne podejście do badań statystycznych, w tym w szczególności instytucji statystyki publicznej, bierze pod uwagę konieczność korzystania z wielu źródeł danych, będącą odpowiedzią na pojawiające się zagrożenia w realizacji badań reprezentacyjnych.

Integracja źródeł danych statystycznych, wraz z danymi administracyjnymi, jest jedną z najważniejszych możliwości utrzymania lub poprawy jakości badań statystycznych. Nowe rozwiązania technologiczne pozwalają ponadto dołączyć do tego zbioru danych informacje mieszczące się w kategoriach *metadata* i *paradata*.

Natomiast *big data* stanowi dla współczesnych badań statystycznych raczej szansę niż zagrożenie. Przede wszystkim możliwości oferowane przez *big data* mogą być wykorzystane w tradycyjnych badaniach statystycznych do zmniejszania wpływu błędów nielosowych na jakość opisu lub wnioskowania. Należy oczekiwać, że oba te podejścia: klasyczne badania statystyczne oraz techniki eksploracji dużych zbiorów danych (*big data*) będą jeszcze przez dłuższy okres czasu komplementarne względem siebie. W niektórych zastosowaniach (np. w predykcji) *big data* wypiera stopniowo badania próbkowe. W tych zagadnieniach z kolei, w których wymagana jest duża precyzja szacunku, wiodące pozostaną badania reprezentacyjne. Rozwój metodyki wnioskowania na podstawie bardzo dużych zbiorów danych powinien uwzględniać zarówno ograniczenia klasycznego wnioskowania statystycznego, jak i (w wielu przypadkach) nieadekwatność założenia o losowości próby. Są to ważne wyzwania, jakie podejmują statystycy, mając na uwadze oczekiwania wyrażane w tym względzie przez badaczy zagadnień ekonomicznych i społecznych.

Bibliografia

- Baker, H.K., Mukherjee T.K. (2007), Survey Research in Finance: Views from Journal Editors, *International Journal of Managerial Finance*, Vol. 3(1), s.11-25.
- Dippo C.S., Sundgren B. (2000), "The Role of Metadata in Statistics", paper presented at the *International Conference on Establishment Surveys II*, Buffalo, New York.
- Donsbach W., Traugott M.W. (ed.) (2008), *Public Opinion Research*, SAGE Publications.
- Goldfarb A., Lu Q. (2006), Household-specific regressions using clickstream data, *Statistical Science*, vol. 21(2), s. 247–255.
- Heneley J. (2011), Do we actually need a census", *The Guardian*, 10.03.2011.
- Paradysz J. (2010), Konieczność estymacji pośredniej w spisach powszechnych, [w:], E. Gołata (red.), *Pomiar i informacja w gospodarce*, „Zeszyty Naukowe" nr 149, Uniwersytet Ekonomiczny w Poznaniu.
- Greene W. (2003), *Econometric Analysis*, (5th ed.), Prentice Hall, New York.
- GUS (2015), Statystyka publiczna – współczesne oblicze, Warszawa.
- Kmenta J. (1990), *Elements of econometrics*, Macmillan Publishing Company, New York.
- Hubbard R., Armstrong J. (2006), Why we don't really know what statistical significance means: A major educational failure, *Journal of Marketing Education*, 28, s. 114–120
- Kreuter F. (2015), "Paradata", [w:] Krosnick J.A., S. Presser, K. Husbands Fealing, S. Ruggles (Eds.), *The Future of Survey Research: Challenges and Opportunities*. Arlington, VA: The National Science Foundation Advisory Committee for the Social, Behavioral and Economic Sciences Subcommittee on Advancing SBE Survey Research.
- 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), s. 906-917.
- Mayer- Schönberger V., Cukier K. (2013), *BIG DATA. Rewolucja, która zmieni nasze myślenie, pracę i życie*. Wyd. MT Biznes, Warszawa.
- Platek R., Särndal C.E. (2001), Can a statistician deliver?, *Wiadomości Statystyczne* nr 4.
- Särndal C.E., LundströmS. (2006), *Estimation in Surveys with Nonresponse*, J. Wiley.
- Stefanowicz B., Cierpiał-Wolan M. (2015), Błędy przetwarzania danych, *Wiadomości Statystyczne*, nr 9, s. 23-29.
- Szreder M. (2007), O roli informacji spoza próby w badaniach sondażowych. *Przegląd Socjologiczny* tom LVI/1, s. 97-108.

- Szreder M. (2010), *Metody i techniki sondażowych badań opinii* (wyd. II zmienione). Polskie Wydawnictwo Ekonomiczne, Warszawa.
- Szreder M. (2013), Twierdzenie Bayesa po 250 latach. *Wiadomości Statystyczne* nr 12, s. 23-26.
- Szreder M. (2015), Zmiany w strukturze całkowitego błędu badania próbkowego. *Wiadomości Statystyczne* nr 1, s. 4-12.
- Szreder M. (2016), O niektórych nowych wyzwaniach i oczekiwaniach wobec statystyki. *Wiadomości Statystyczne* nr 6, s. 1-9.
- Wallgren A., Wallgren B. (2007), *Register-based Statistics. Administrative Data for Statistical Purposes.* John Wiley, New York.

Streszczenie

W artykule scharakteryzowano najważniejsze tendencje w zakresie rozwoju badań statystycznych i ich zastosowań w naukach ekonomicznych. Szczególną uwagę zwrócono na konsekwencje wynikające z postępu technologicznego, ułatwiającego dostęp do wielu źródeł danych, a także zwiększającego możliwości pozyskiwania, przetwarzania i analizy danych statystycznych. Podkreślono, że zwiększające się znaczenie błędów nielosowych w badaniach reprezentacyjnych, powoduje konieczność korzystania ze zintegrowanych zbiorów danych oraz z informacji określanych terminami *paradata* i *metadata*. Analizując natomiast najnowsze trendy w badaniach statystycznych, wskazano zarówno na komplementarną rolę *big data* w tradycyjnych badaniach statystycznych (w szczególności niewyczerpujących), jak i na stopniowe zastępowanie w niektórych zagadnieniach badań próbkowych analityką *big data*.

Słowa kluczowe: badania reprezentacyjne, błędy nielosowe, big data

Abstract

The paper discusses some major trends in the development of statistical surveys and in their applications in economics. It focuses on consequences of the technological revolution which facilitates access to various data sources, and creates new opportunities of collecting, processing and analyzing data. The author justifies that an increase in the impact of non-random errors in sample surveys implies the necessity of using a combination of available data sources, including paradata and metadata. With regard to big data, two tendencies have been highlighted: complimentary role of big data in traditional statistical surveys, and – in some areas of research – replacing sample surveys by big data analyses.

Key words: sample surveys, non-random errors, big data.

Notka biograficzna autora

Prof. dr hab. Mirosław Szreder jest statystykiem zajmującym się zagadnieniami klasycznego i bayesowskiego wnioskowania statystycznego, a także teorią i praktyką badań reprezentacyjnych. Jest m.in. autorem książki pt. „Metody i techniki sondażowych badań opinii". Znany jest także jako popularyzator nauki – autor artykułów na temat rozwoju i zastosowań statystyki, publikowanych w dziennikach i tygodnikach społeczno-politycznych.