

OD TWIERDZENIA JAKUBA BERNOULIEGO DO WSPÓŁCZESNYCH BADAŃ REPREZENTACYJNYCH

Jan Kordos

Wyższa Szkoła Menedżerska w Warszawie

- **Bernoulli, J. (1713), *Ars Conjectandi* (Sztuka przewidywania), , Basilea: Thurnisius**
- **Bayes, T. (1763), "An essay towards solving a problem in the doctrine of chances," *Philosophical Transactions of the Royal Society*, 53, 370–418.**
- **Neyman J. (1933), *Zarys teorii i praktyki badania struktury ludności metodą reprezentacyjną*, Instytut Gospodarstwa Społecznego, Warszawa, 1933.**

(Podróż przez trzy stulecia)

- **Kish, L. (1996), Stulecie zmagania o badania reprezentacyjne, *Wiadomości Statystyczne*, nr 8, s. 3-16.**
- **Kuusela, V. (2011), „*Paradigms in Statistical Inference for Finite Populations Up to the 1950s*, Research Reports 257, © 2011 – Statistics Finland**

Jacob Bernoulli

Ars Conjectandi

1713

JACOBI BERNOULLI,
Profess. Basil. & utraque Societ. Reg. Scientiar.
Gall. & Pruss. Sodal.
MATHEMATICIS CELEBRISSIMUS,
ARS CONJECTANDI,
OPUS POSTHUMUM.
Accedit
TRACTATUS
DE SERIEBUS INFINITIS,
Et Epistola Gallicè scripta
DE LUDO PILÆ
RETICULARIS.



BASILEÆ.
Impensis THURNISIORUM, Fratrum.
cdo lxxx xlii.

PLAN PREZENTACJI

1. Twierdzenie Jakuba Bernoulliego, prawo wielkich liczb i centralne twierdzenie graniczne.
2. Przejście z teorii do praktyki → tablice liczb losowych
3. Wykorzystanie liczb losowych w praktyce
4. Metodologiczne problemy badań reprezentacyjnych:
 1. Metody wyboru próby:
 2. Metody estymacji
 3. Metody zbierania danych
5. Rozwój badań reprezentacyjnych w ostatnich latach
6. Zaniedbane dziedziny badań reprezentacyjnych.
7. Szacowanie ogólnego błędu badania.
8. Wnioski końcowe

Twierdzenie Jakuba Bernoulliego

S. Zubrzyckiego (Zubrzycki, 1966, s. 139):

Jeśli $X_1, X_2, \dots, X_n$ jest ciągiem niezależnych zmiennych losowych o jednakowym rozkładzie prawdopodobieństwa z wartością oczekiwaną μ i odchyleniem standardowym σ , a $\bar{X}$ oznacza średnią pierwszych n zmiennych losowych, to dla każdego $\varepsilon > 0$ i $\eta > 0$ zachodzi nierówność:

$$P(|\bar{X} - \mu| < \varepsilon) > 1 - \eta$$

gdy n jest dostatecznie duże.

Teoria i praktyka:

TABLICE LICZB LOSOWYCH

„Jeśli $X_1, X_2, \dots, X_n$, jest ciągiem **niezależnych** zmiennych losowych o jednakowym rozkładzie prawdopodobieństwa z wartością oczekiwaną μ i odchyleniem standardowym σ ”.

Aby spełnić ten warunek teoretyczny w praktyce,
statystycy zbudowali:

Tablice liczb losowych

Druga część twierdzenia słabego prawa wielkich liczb brzmi:

„wtedy dla każdego $\varepsilon > 0$ i $0 < \eta < 1$ zachodzi ,

$$P(|\bar{X} - \mu| < \varepsilon) > 1 - \eta$$

gdy n jest dostatecznie duże”

Wydaje mi się, że prof. J. Neyman przy tworzeniu przedziałów ufności wykorzystał nierówność $P(|\bar{X} - \mu| < \varepsilon) > 1 - \eta$

i zastąpił ją , $P(|\bar{X} - \mu| < \varepsilon) = 1 - \alpha$

gdzie ε oznacza precyzję oceny, a $1 - \alpha$ poziom ufności.

LICZBY LOSOWE I ICH ZASTOSOWANIA

Losowanie i liczby losowe wykorzystywano już w XVIII wieku i później :

- w ruletkach,
- badaniach eksperymentalnych, a także
- w badaniach częściowych na mniejszą skalę.

Pierwsze tablice liczb losowych wydał w roku 1927

L. H. Tippett pod tytułem „*Random Sampling Numbers*”.

- Uwagi K. Persona w przedmowie do tablic Tippetta
- Liczby losowe wykorzystane są w reprezentacyjnych badaniach statystycznych (przy losowaniu próby z populacji generalnej lub, szerzej, planowaniu schematu losowania), a zatem,
 - *badaniach ekonomicznych,*
 - *społecznych,*
 - *marketingowych,*
 - *badaniach opinii*
 - *w statystycznej kontroli jakości produktów, itp.*

Innym sposobem wykorzystania liczb losowych są wszelkie badania symulacyjne.

- Zastosowaniem liczb losowych, które w ostatnim czasie zyskało ogromnie na znaczeniu, jest *kryptografia*, czy też szerzej rozumiana *ochrona informacji*.

(Kotulski, 2001; Zieliński, 1970, 1972, 1992).

PRAWO WIELKICH LICZB
CENTRALNE TWIERDZENIE GRANICZNE

PRAWO WIELKICH LICZB

B. Pascal (1623–1662) i C. Huygens (1629–1695).
J. Bernoulli (1654–1705)
S. D. Poisson (1781–1840), P.L. Czebyszew (1821–1894).

CENTRALNE TWIERDZENIE GRANICZNE

de Moivre-Laplace,
Levy, Lindeberg, Lapunow, Feller i Chinczyn

Odkrywczy rozkładu normalnego:

A. de Moivre (1733,), C.F. Gauss (1809),
B. A. Laplace (1812)

METODOLOGICZNE PROBLEMY BADAŃ REPREZENTACYJNYCH

Metodologicznie, badania reprezentacyjne dotyczą trzech różnych, lecz równie ważnych pytań:

- 1. W jaki sposób próba powinna być wybrana z danej populacji (metody wyboru próby).***
- 2. W jaki sposób wyciągnąć wnioski o populacji w oparciu o dane otrzymane z wybranej próby oraz dostępnych informacji (metody estymacji), oraz***
- 3. W jaki sposób uzyskać informacje od wybranych jednostek do próby i dostępnych źródeł (metody pozyskiwania danych).***

METODY WYBORU PRÓBY

Wpływ Bernoulliego:

- Losowy wybór próby
- Badania monograficzne
- Badania eksperymentalne
- Metody wyboru celowego

Powstanie metody reprezentacyjnej:

- Początki metody reprezentacyjnej (1895)
- Dyskusja na sesjach Międzynarodowego Instytutu Statystycznego (MIS)
- Uznanie metody reprezentacyjnej przez MIS w 1925 r.

Badania reprezentacyjne na początku XX w.

- wybór losowy
- Wybór celowy

Wpływ Jerzego Neymana

- Artykuły (1934, 1938)
- Hansen i in. (Biuro Spisów USA)

Nowoczesne metody wyboru próby

- Oddzielne przedstawienie

METODY ESTYMACJI

1. Wpływ Bernoulliego

- Wykorzystane twierdż. Bernoulliego przez Laplace'a do uogólnienia wyników z próby na całą populację.
- Wpływ de Moivre'a w rozpowszechnieniu pracy Bernoulliego.

2. Wpływ Bowleya :

- o pomiarze (Bowley, 1915) oraz
- o doborze próby (Bowley, 1926)

3. Fundamentalny wkład Jerzego Neymana:

- Neyman J. (1934), On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection, *"Journal of the Royal Statistical Society"*, nr 97, s. 558-606.
- Neyman, J. (1938), Contribution to the Theory of Sampling Human Population, *Journal of American Statistical Association*, nr 33, s.101-116.

4. Złożone metody estymacji

- Hansen i in.; Biuro Spisów USA

PRZEŁOM W METODZIE REPREZENTACYJNEJ

W pierwszych dekadach XX wieku podejmowane są już badania częściowe na szerszą skalę

1. Należy wspomnieć o badaniach prowadzonych i opisanych przez A.L. Bowleya'a (1913, 1926) w Anglii, a także
2. badania bezrobocia prowadzone w latach 1934-1942 w USA (Frankel i Stock, 1942; Stephan, 1948), które w 1943 r. przekształciły się w sławne badania siły roboczej prowadzone przez Biuro Spisów USA.

Międzynarodowy Instytut Statystyczny (MIS) uznał metodę reprezentacyjną w statystyce oficjalnej dopiero w 1925 r.,

Fundamentalny wkład w rozwój teorii metody reprezentacyjnej przez *Jerzego Neymana (1934)*

- zmienił teoretyczne podstawy wnioskowania na podstawie badań częściowych, wprowadzając błędy losowe oparte na randomizacji rozkładu;
- Wprowadził *optymalną lokalizację próby* w losowaniu warstwowym, a także *przedziały ufności*;
- *Opracował teorie losowania podwójnego z populacji skończonej (zastosowanie w praktyce).*

ZŁOŻONE METODY ESTYMACJI

Estymacja Bayesowska

» Empiryczna

» Hierarchiczna

(Dla małych obszarów)

Szacowanie wariancji

- Metody grup losowych
- Linearyzacja Taylora
- Met. Jackknife
- Powtórne replikacje bilansowe
- Metody bootstapowe

Imputacja

(Brak odpowiedzi – różne metody)

Kalibracja

- Kalibracja stała się ważnym narzędziem metodologicznym w badaniach reprezentacyjnych na dużą skalę
- Niektóre krajowe urzędy statystyczne opracowały oprogramowanie przeznaczone do obliczenia wag.

Deville J-C., Särndal C-E. (1992), „Calibration Estimators in Survey Sampling”, *Journal of the American Statistical Association*, Vol. 87, 376–382.

Särndal C-E. (2007), „The Calibration Approach in Survey Theory and Practice”, *Survey Methodology*, Vol. 33, No. 2, 99–119.

ROZWÓJ BADAŃ REPREZENTACYJNYCH W DRUGIEJ POŁOWIE XX WIEKU

Po 1945 r. następował lawinowy rozwój badań reprezentacyjnych w wielu krajach.

1. Ukazało się 5 podręczników z metody reprezentacyjnej w języku angielskim: Yates (1949), Deming (1950), Chochran (1953), Hansen, Hurwith i Madow (1953) oraz Sukhatme (1954).
2. Po ukazaniu się wymienionych podręczników zorganizowano wiele kursów z metody reprezentacyjnej w USA, Wielkiej Brytanii, w Indiach i wielu innych krajach.
3. W USA badania prowadzone przez Biuro Spisów przekształciło się w Bieżące Badania Ludności (Current Population Surveys), które pod kierunkiem M.H Hansena wykształciło wielu specjalistów w dziedzinie badań reprezentacyjnych na poziomie krajowym i międzynarodowym.
4. W latach 1940 -1960 szczególną rolę w badaniach reprezentacyjnych odegrali Hansen, Hurwitz, Madow, Waksberg, Bershad, Tepping i inni.
 - *Uważa się, że Morris Hansen odegrał szczególną rolę w rozwoju metodologii badań reprezentacyjnych w XX wieku.*
 - *Hansen et al. (1943) zastosowali losowanie warstwowe wielostopniowe z jedną jednostką pierwszego stopnia w warstwie z prawdopodobieństwem proporcjonalnym do wielkości, aby zapewnić zrównoważony schemat losowania*

ROZWÓJ BADAŃ REPREZENTACYJNYCH W DRUGIEJ POŁOWIE XX WIEKU (cd.)

Instytut Statystyczny w Indiach,
kierowany przez Mahalanobisa (1944, 1952),
stał się drugim centrum badań reprezentacyjnych
o znaczeniu międzynarodowym.

- *Indyjskie Narodowe Badanie Reprezentacyjne (Indian National Sample Surveys) zdobyło sławę międzynarodową (Suthatme i Sukhatme, 1970).*
- *Mahalanobis (1944) już w 1937 r. w Indiach wykorzystał losowanie wielostopniowe (wsi, poletek i zdjęć fotograficznych) do badania plonów.*
- *Wykorzystał funkcje kosztu badania i wariancje przy planowaniu badania.*

Koncentracja na błędach nielosowych

- *W latach 1950-tych prawie zakończono rozwój teorii randomizacji, a zainteresowanie statystyków przesunęło się stopniowo z błędów losowych w kierunku błędów nielosowych,*
- *W szczególnym stopniu przyczynił się M.H. Hansen i jego współpracownicy (Hansen i Hurwitz, 1943, 1946; Hansen, Hurwitz i Madow, 1953; Hansen i Steinberg, 1956; Hansen, Hurwitz i Bershad, 1961; Hansen i Waksberg, 1970).*

ROZWÓJ BADAŃ REPREZENTACYJNYCH W DRUGIEJ POŁOWIE XX WIEKU (cd.)

W aspekcie międzynarodowym znaczną rolę odgrywają także sesje naukowe *Międzynarodowego Instytutu Statystycznego*, które odbywają się co dwa lata.

1. *Na sesjach tych prezentowane są zarówno osiągnięcia teoretyczne jak i praktyczne w tej dziedzinie uzyskane przez ośrodki badawcze różnych krajów, a także przez urzędy statystyczne.*
2. *Istotną rolę w rozwoju metodologii badań reprezentacyjnych odgrywa także jedna z sekcji MIS: International Association of Survey Statisticians (IASS), która jest organizatorem lub współorganizatorem wielu konferencji międzynarodowych poświęconych różnym aspektom badań reprezentacyjnych.*
3. *W Polsce organizacja ta była współorganizatorem międzynarodowej konferencji poświęconej metodom estymacji dla małych obszarów w 1992 r. (Kalton et al. 1993).*

Dość zróżnicowany rozwój badań reprezentacyjnych był w ośrodkach uniwersyteckich.

(Warto tu wspomnieć o dwóch ośrodkach):

- *Centrum Badań Reprezentacyjnych (Sample Surveys Center) na uniwersytecie Iowa (Iowa State University), zorganizowane przez Snedecora i Cochran (1977), oraz*
- *Centrum Badawcze na Uniwersytecie Michigan (Survey Research Center) kierowane przez L. Kisha (Kish, 1965, 1996, 1987).*

METODY POZYSKIWANIA DANYCH

Klasyczne metody:

- wywiadu,
- obserwacji,
- Prowadzenie zapisów
- wykorzystanie innych źródeł.

Wykorzystanie nowoczesnych technik obliczeniowych:

- **Wywiady wspomagane komputerem** - CAI (Computer Assisted Interview),
- **Wywiady telefoniczne:** CATI (Computer Assisted Telephone Interview)
- **Wykorzystanie Internetu; Spis Ludności i Mieszkań 2011**
- **Wykorzystanie dostępnych źródeł administracyjnych i finansowych.**

KTÓRE MODELE STATYSTYCZNE SĄ UŻYTECZNE?

- W niektórych praktycznych zastosowaniach w szerokim zakresie wykorzystane są różnego rodzaju modele przy estymacji wyników i ich analizie.
- Powstaje pytanie, czy zawsze możemy wykorzystywać dostępne modele statystyczne?
- Nawiążę tu do słynnego stwierdzenia Boxa iż „*wszystkie modele są złe, ale niektóre są użyteczne*” („*All models are wrong, but some are useful*”).
[G.E.P. Box (1976) [*Science and Statistics*](#) *Journal of the American Statistical Association*, Vol. 71, No. 356., s. 791-799.]
- Często w badaniach reprezentacyjnych uzyskuje się oceny bardzo precyzyjne a zarazem niedokładne, np. z zakresu dochodów, w których obciążenie nieraz przewyższa 10%, a precyzja kształtuje się poniżej 2%.
- W takich przypadkach należy wykorzystywać tylko precyzję, a stosowanie przedziałów ufności jest błędne.

Ocena błędu całkowitego w badaniach reprezentacyjnych

Kish, L. (1996), Stuletnie zmagania o badania reprezentacyjne, *Wiadomości Statystyczne*, nr 8, s. 3-16.

- Smith, T.M.F. (1976). The foundations of Survey Sampling: a review. *Journal of Royal Statistical Society*, A. 139, s. 183-204.
- Smith, T.M.F. (1994). Sample Surveys 1975-1990; an age of reconciliation? *International Statistical Review*, 62, s. 3-34.

Dla niektórych powolny postęp w modelowaniu całkowitego błędu badania jest zniechęcający. Smith (1990) zauważa, że „...*potrzebny jest systematyczny plan redukowania błędu całkowitego, jeśli mamy mieć na uwadze w badaniach ogólne zarządzanie jakością*”.

Ten sam autor, kilka lat później (1994) zauważył: „*Porażka zarówno statystyków teoretyków jak i praktyków w zintegrowaniu błędów losowych i nielosowych w wyrażeniu całkowitego błędu badania nawet po 50 latach intensywnych prac badawczych, musi być uważana jako jedna z bardziej dokuczliwych porażek tej ważnej branży statystyki*”.

Linacre, S.J. i Trewin, D.J. (1989). Evaluation of errors and appropriate resource allocation in economic collection. Australian Bureau of Statistics, *Statistical Services Papers*, vol. 7, 1, s. 33-50.

WNIOSKI KOŃCOWE

PRACE EUROSTATU W ZAKRESIE JAKOŚCI STATYSTYKI

- Eurostat (2007), *Handbook on Data Quality Assessment: Methods and Tools*.
- Eurostat (2009), *Handbook on Quality Reports*.
J. Bernoulli
- Eurostat (2013), *Handbook on improving quality by analysis of process variables*.