University of Gdańsk

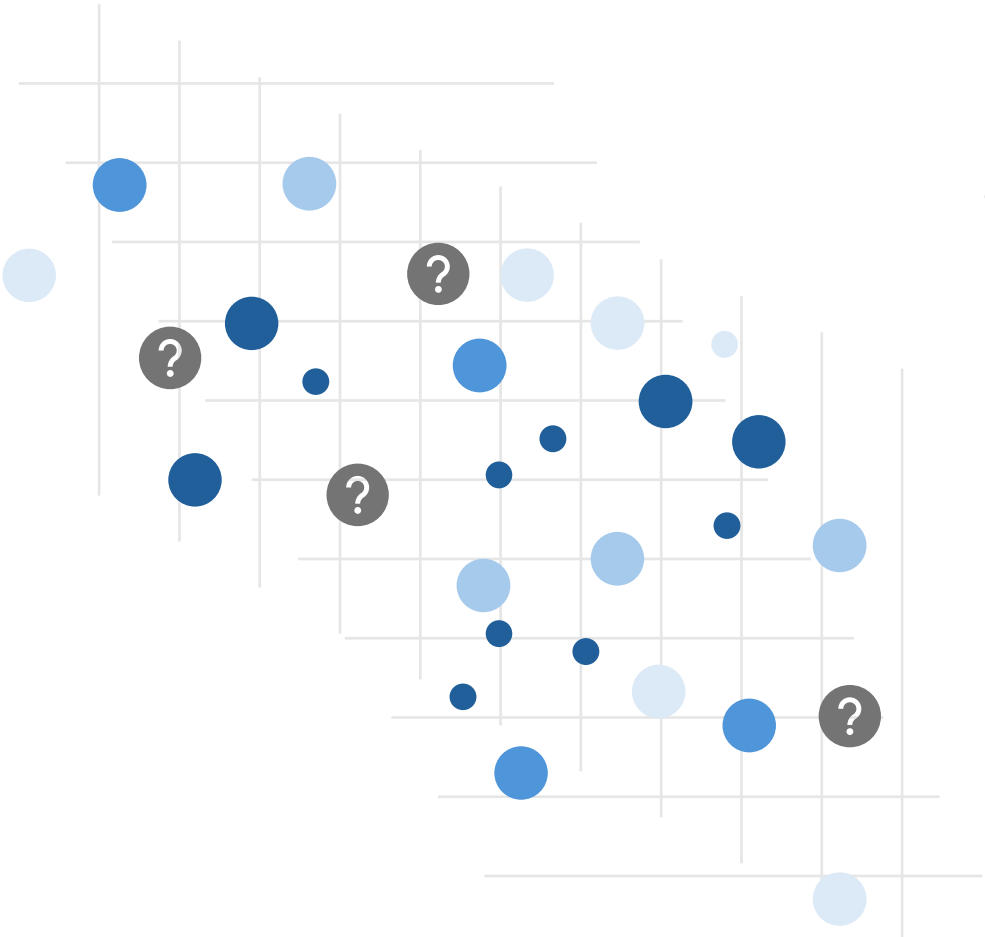
Mirosław Szreder

Representative sample – need and proposal for a definition

The 4th Congress of Polish Statistics:
July 2-4, 2024, Warsaw, Poland

Motto

Statistics is a guide to the unknown



(Robert V. “Bob” Hogg,
Professor Emeritus of Statistics
and Actuarial Science, 1924-2014)

Source: David S. Moore (1998).
Statistics among the Liberal Arts,
„Journal of the American
Statistical Association”,
93:444, pp. 1253-1259.

Quality of Inference

Quality of statistical inference is determined mainly by the following factors:

1) the kind (nature) of the sample, the way it is selected
   $\longrightarrow$ **representativeness**;
2) an opportunity to minimize or to avoid nonrandom errors;
3) prior information about the population and ways it is incorporated;
4) the sample size;
5) method of data collection and mode of administration the survey (CATI, CAWI,...).

Sample size and its influence on the quality of inference

Too often it is assumed that the overall quality of inference will improve if the sample size increases.

Some people equate **the total survey error** with **the random error**.

Only the latter one is a decreasing function of the sample size.

Sometimes nonrandom errors are most challenging regardless of the sample size. They involve:

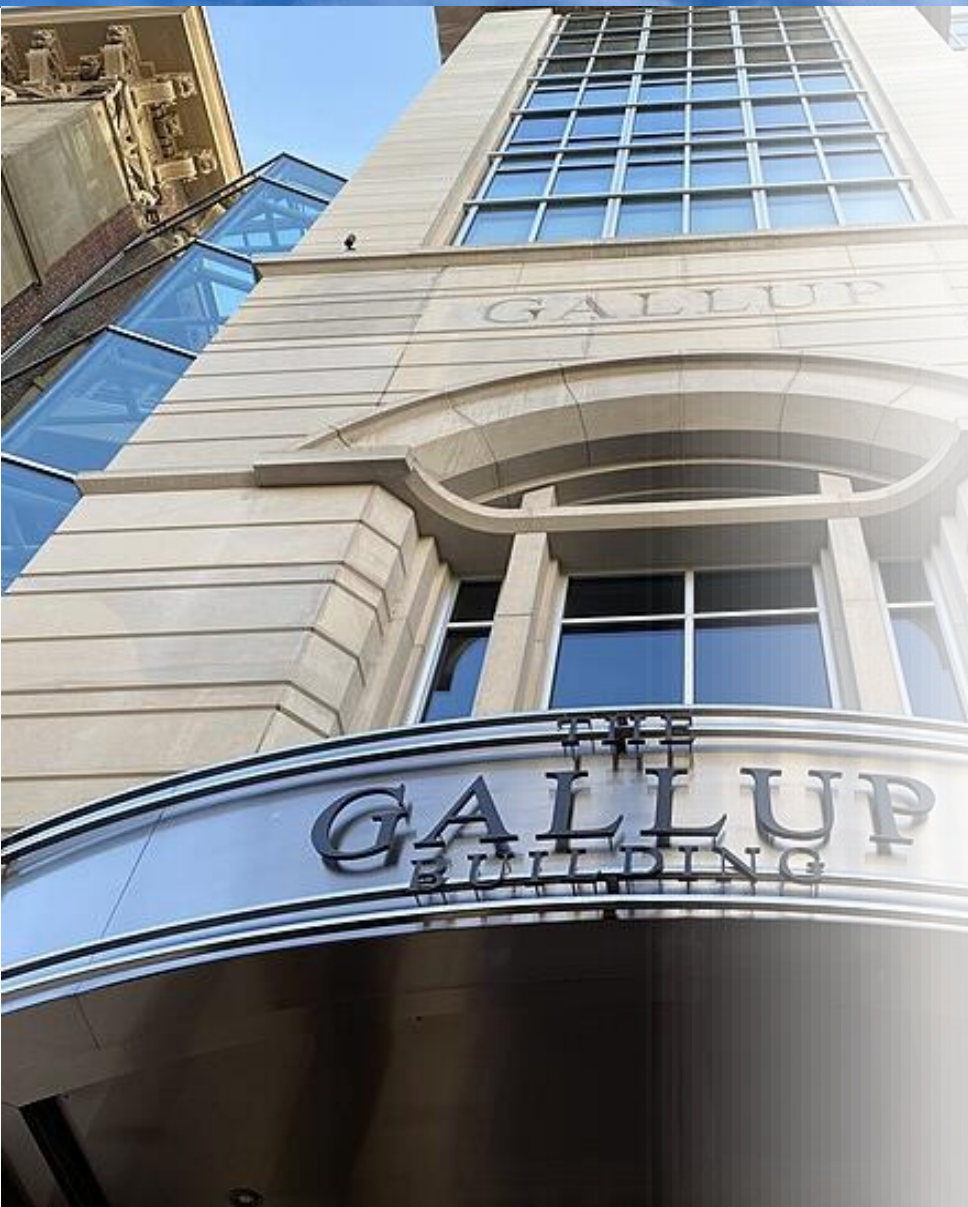
- coverage error
- nonresponse error
- measurement error
- data processing error.

Sample size and its influence on the quality of inference

The following statements are **not** true:

- ✓ large sample sizes guarantee good quality of inference,
- ✓ small samples (minor fractions of the population) cannot be regarded as a good basis for the high quality inference.

Sample size and its influence on the quality of inference



Gallup interviews **a minimum of 1,000 U.S. adults** aged 18 and older for each Gallup Poll Social Series survey.

Results for this Gallup poll are based on telephone interviews conducted Nov. 2-8, 2017, with **a random sample of 1,028 adults**, aged 18 and older, living in all 50 U.S. states and the District of Columbia. For results based on the total sample of national adults, the margin of sampling error is ±4 percentage points at the 95% confidence level.

Sample size and its influence on the quality of inference

Exit poll

The number of respondents less than 1%.

Parliamentary Elections in Poland, 15 October, 2023:

	Exit poll 9 p.m.	Final (official) result	Error (p.p.)
PiS	36.8	35.4	1.4
KO	31.6	30.7	0.9
Third Way	13.0	14.4	1.4
The Left	8.6	8.6	0.0
Confederation	6.2	7.2	1.0
Local governments (politically unaffiliated)	2.4	1.9	0.5
Other	1.2	1.6	0.4

Sample size and its influence on the quality of inference

Exit poll

The number of respondents less than 1%.

The 2024 European Parliamentary election (Poland):

	Exit poll	Final (official) result	Error (p.p.)
PiS	33.9	36.2	2.3
KO	38.2	37.1	1.1
Confederation	11.9	12.1	0.2
Third Way	8.2	6.9	1.3
The Left	6.6	6.3	0.3

Sample size and its influence on the quality of inference

Exit poll

2022 Parliamentary Elections in Italy:

- Exit poll prediction of Giorgia Meloni right wing party result: 41% - 45%
- Final result: 43.7%
- The mid-point of the exit poll interval assessment is 43%

  $\longrightarrow$ the error turned out to be less than 1%.

Sample size and its influence on the quality of inference

Exit poll

In Great Britain the exit poll forecast usually relates to the number of seats for political parties in the Parliament

- Year 2005: the errors was less than 0.8%
- Year 2010: the estimated number of seats for three main parties (table below). The maximal error did not exceed 3 seats.

The 2010 exit poll (commissioned jointly by BBC, ITV and Sky)				
	Con	Lab	LD	Other
Prediction at 10pm:	307	255	59	29
Actual seats won:	307	258	57	28

Mode of administration the survey does not determine the sample character

The IBRiS survey for Polsat's "News" was conducted on May 19-21, among a group of 1,020 people. The survey was carried out using the CATI method - telephone, standard, computer-assisted questionnaire interviews.

DGP Dziennik Gazeta Prawna (May 22, 2023)

The "We care about the Poles" survey conducted by Kantar Millward Brown on behalf of *Wysokie Obcasy* was conducted using the CATI method on March 5-11, 2019, based on the declarations of 1,500 people representing the population of adults 18+.

EuroPAP News (https://europapnews.pap.pl/) from 13.03.2019 r.

From October 27 to November 7, 2014, the INDICATOR Marketing Research Center conducted a survey of employers' preferences in terms of employing MBA graduates. The study was conducted using the CATI (Computer Assisted Telephone Interviews) method. The measurement was carried out on a sample of $N = 450$ enterprises employing more than 20 employees under an employment contract.

„Perspektywy" (https://perspektywy.pl)

Representative sample in scientific inference and statistical inference

Rothman K.J., Gallacher J.E., and Hatch E.E., ***Why representativeness should be avoided***, „International Journal of Epidemiology", 2013, t.42, no 4, pp. 1012–1014.

Richiardi L., Pizzi C., and Pearce N., *Commentary:* ***Representativeness is usually not necessary and often should be avoided***, „International Journal of Epidemiology", 2013, t.42, no 4, pp. 1018–1022.

Ebrahim S., and Davey Smith G., *Commentary:* ***Should we always deliberately be non-representative?***, „International Journal of Epidemiology", 2013, t.42, no 4, pp. 1022–1026.

Representative sample in scientific inference and statistical inference

Scientific Inference

vs.

Statistical Inference

Exploratory objectives of scientific research - striving to formulate universal laws of nature, not confined to a specific place and fixed time of observation; aiming to learn the causes and mechanisms of actions, raise new questions and improve methods for further research.

Quote:

An exploratory study explores new questions rather than tests an existing hypothesis. But scientists have felt that they had to disguise an exploratory study as hypothesis testing and that is totally dishonest.

(Marcia McNutt, president of the U.S. National Academy of Sciences).

Representative sample in scientific inference and statistical inference

Scientific Inference

vs.

Statistical Inference

Confirmatory and descriptive objectives of statistical research – focused on testing hypotheses and generalizations on the specific population (determined by time and place).

Box G.E.P., *Statistics as a Catalyst to Learning by Scientific Method Part II—A Discussion*, „Journal of Quality Technology", 1999, t.31, no 1, pp. 16–29.

Statistics has no reason for existence except as a catalyst for scientific enquiry in which only the last stage, when all the creative work has already been done, is concerned with a final fixed model and a rigorous test of conclusions.

(George E.P. Box)

Representative sample in scientific inference and statistical inference

While sample representativeness is considered a *sine qua non* condition for the credibility of social and economic surveys, in medical and natural sciences, this category has a different status.

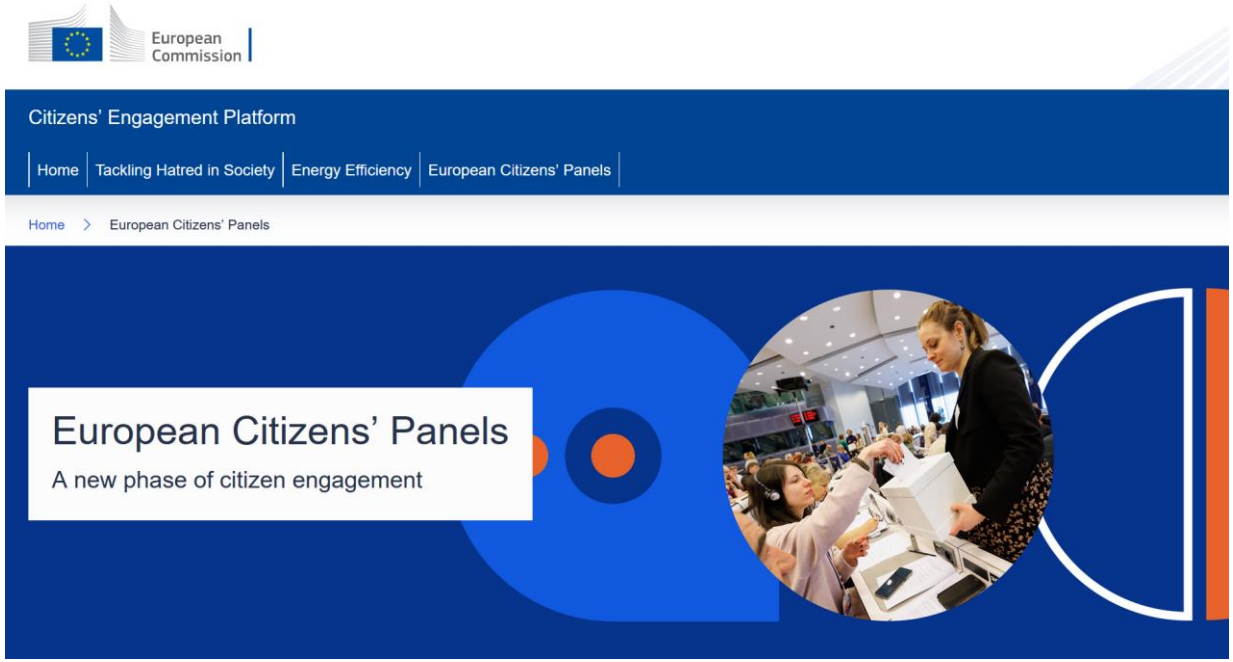
Biologists, immunologists, … select highly non-representative, homogeneous samples, which consist of individuals having identical genes, living in the same environment and using an identical diet.

In experimental sciences, a distinction is sometimes made between generalisability of estimate and generalisability of interpretation.

It also applies to medicine, except for descriptive evaluations, e.g. a given disease proliferation.

Representativeness versus randomness

Citizens' Panels are becoming a regular feature of democratic life in the European Union (EU). They bring together **randomly-selected citizens** from all 27 member states to discuss – at European level – key, upcoming proposals that affect us all.



European Citizens' Panels see participants working together in small groups (each of around 12 people) and all together (in plenaries). A facilitation team provides support. Based on the discussions, citizens make recommendations for the European Commission to consider when defining policies and initiatives.

Representativeness versus randomness

Statistical inference is inherently connected with random sampling for two reasons:

1) a random sample reflects well, on average, the actual structure of the population;

2) it enables the statistician to state the probable accuracy of his/her results within clearly defined limits.

The role of prior knowledge

Prior knowledge about the target population seems to be the key source of information in ensuring that the sample structure is as similar as possible to the structure of the population.

Quota sampling

Stratified sampling

Proposed definition of a representative sample

“A representative sample is one that ensures external validity in relationship to the population of interest the sample is meant to represent”.

(Lavrakas P.J. (2008). Encyclopedia of survey research methods, vol. 2. SAGE Publications, California)

“We define a study sample to be representative of a well-defined target population if the results estimated in that sample are generalizable to the target population”.

(Rudolph J.E., Zhong Y., Duggal P., Mehta S., Lau B. (2023). Defining representativeness of study samples in medical and population health research. “BMJ Medicine”).

Authors do not emphasize problems of errors and their measurement.

Proposed definition of a representative sample

Proposed definition:

A representative sample is one whose composition is identical to or close to the composition of the population, and which can constitute the basis for generalizations of the results to the population with assigned errors.

Apart from this, there are samples whose structures are identical to the population structure with respect to only some selected characteristics.

These should be called representative in relation to the specific control characteristics, and not representative in general (according to the definition proposed above).

University of Gdańsk

Mirosław Szreder

Thank you for your attention

The 4th Congress of Polish Statistics:
July 2-4, 2024, Warsaw, Poland