INTRODUCTION TO SAMPLE AND SAMPLING IN BIOMEDICINE

Lampret Puž, Stella

World of Health 2024; (7) : 43-47 ISSN 2623-5773

PREGLEDNI RAD

INTRODUCTION TO SAMPLE AND SAMPLING IN BIOMEDICINE

UVOD U UZORKE I UZORKOVANJE U BIOMEDICINI

Stella Lampret Puž *[1]

[1] *Fakultet zdravstvenih studija Sveučilišta u Rijeci, Katedra za temeljne medicinske znanosti, Rijeka, Hrvatska*

**Autor za korespondenciju:*
stella.lampret.puz@fzsri.uniri.hr
ORCID: 0009-0000-3799-1707

ABSTRACT

In scientific research, the population is the largest statistical group and consists of all the studied subjects placed in a particular time and place. Subjects express variability, and the two samples can never be equal but only similar. It is not possible to use the entire population due to limited time for research, financial issues, and, most importantly, the real availability of subjects. Hence, scientists developed a research methodology based on a sample, part of the population, generated through the process of sampling. Sampling produces a sample, but a sampling error is also described in statistics by standard error. The most important sample feature is representativeness, achieved by adequate calculation of sample size and random selection of subjects from the population that differentiates two sample types: probabilistic (random selection) and nonprobabilistic (not representative samples). Probabilistic sample types include simple random samples, stratified samples, systematic samples, and cluster samples, while nonprobabilistic samples include convenience samples and quota samples. The sample size is the only factor that the researcher can control while planning the study. However, for a sample to be representative, it is necessary to define inclusion and exclusion criteria to ensure that the respondents included in the sample have characteristics that affect the association being researched. It directly affects the statistical power of the study, and therefore, sample size has to be calculated using the statistical methods described in this paper.

Keywords: *; calculation, population; sample; sampling; standard error*

SAŽETAK

U znanstvenim istraživanjima populacija predstavlja najveći statistički skup koji čine sve jedinke istraživanog sustava obuhvaćene u određeno vrijeme i na određenom mjestu. Jedinke izražavaju varijabilnost, a dva uzorka nikada neće biti jednaka, već samo slična. Nemoguće je uključiti cijelu populaciju u uzorak zbog vremenskog ograničenja istraživanja, financijskih pitanja te, najvažnije, dostupnosti jedinki. Stoga su znanstvenici razvili metodologiju istraživanja temeljenu na uzorku, dijelu populacije koja nastaje procesom uzorkovanja. Uzorkovanjem se dobiva uzorak, ali se pogrješka uzorkovanja (odstupanje uzorka od populacije) u statistici opisuje standardnom pogreškom. Najvažnija značajka uzorka je reprezentativnost koja se postiže odgovarajućim izračunom veličine uzorka i slučajnim odabirom jedinki iz populacije. Razlikuju se dvije vrste uzoraka – probabilistički (slučajni odabir) i neprobabilistički (nereprezentativni uzorci). Probabilistički uzorci uključuju jednostavne slučajne uzorke, stratificirane uzorke, sustavne uzorke i klaster uzorke, dok neprobabilistički uzorci uključuju uzorke pogodnosti i kvotne uzorke. Veličina uzorka jedini je čimbenik koji istraživač može kontrolirati prilikom planiranja istraživanja. Međutim, da bi uzorak bio reprezentativan, potrebno je definirati kriterije uključenja i isključenja kako bi se osiguralo da ispitanici uključeni u uzorak imaju značajke koje utječu na povezanost koja se istražuje. To izravno utječe na statističku snagu istraživanja, stoga se veličina uzorka treba izračunati pomoću statističkih metoda opisanih u ovom radu.

Ključne riječi*: izračun, populacija; uzorak; uzorkovanje; standardna pogreška;*

INTRODUCTION

Population is the largest possible set of elements under observation consisting of all specimens or subjects with certain characteristics (e.g., all patients with some disease or all patients receiving specific therapy) (1,2,3). Population is explained at a particular time and place: it is defined through conceptual, spatial, and temporal definition, ascribing the set of individuals studied, the area where they are located, and the condition of the statistics group, which is limited to one point of the time or a given time span (4). For example, a sample from the published study (3) consists of 35 patients (conceptual definition) with familial Mediterranean Fever collected during 2009 and 2010 years (temporal definition) from the Department of Physical Medicine and Rehabilitation, Ataturk University, Medical Faculty Erzurm, Turkey (spatial definition).

Intervariability is inherent to each subject within a population, and individuals are never equal (identical), only similar (2). Therefore, we cannot take one (any) subject from the population to be the representative of all others. In addition, subject characteristics can change over different time intervals, minutes, hours, or days, and the time of measurement must be noted for certain variables (e.g., blood glucose concentration varies over the day, and female sexual hormones follow specific concentration curves over the month). However, in real life, it is almost impossible to explore an entire population (due to limited research time, financial issues, and the actual availability of subjects, etc.). Therefore, scien-

tists have developed a research methodology based on a part of the population, sample (4).

As the sample represents the whole population, all statistical parameters obtained from the sample correspond to the values of the population parameters (e.g., $\bar{X}$ the sample arithmetic mean represents the arithmetic mean of the population, and the correlation of two numerical variables in the sample represents their correlation in the population). This is why they are assigned by different symbols in statistics (e.g., $\bar{X}$ stands for the sample and μ for the population arithmetic mean, r for the sample and ρ for the population correlation coefficient, p for the sample and π for the population proportion) (5). In fact, as the population cannot be comprehended by research, population parameters are mostly estimated using the values of sample parameters.

REPRESENTATIVENESS

Representativeness is the most important feature of the samples because it implies that all subjects from the sample share the same statistical characteristics of the basic set, population (2). Sample statistical parameters (central tendency and dispersion measurements as, for example, mean and standard deviation) should have equal (or similar) values to those of the same population parameters. If not, the sample is not representative. Good representativeness of the sample to the population is achieved through an adequate size of the sample and the random selection of subjects (5).

Sample representativeness is a condition sine qua non; it must be fulfilled in order to properly conduct the research (6). However, apart from its representativeness, the sample must be measurable and achievable in reality, i.e., subjects or materials from the study can be gathered and monitored, and chosen characteristics can be measured. Furthermore, the representativeness of the sample can be affected by inclusion and exclusion criteria. The mentioned criteria should be clearly defined in order to ensure the inclusion of only those respondents who have characteristics relevant to participation in the research (2).

SAMPLE TYPES

Two basic sample types can be recognized: probabilistic and nonprobabilistic. If each subject from the population has a certain chance of being selected for the sample, a probabilistic sample is achieved. For example, if any blood specimen from the daily routine load in a clinical laboratory might be chosen for quality retesting sample, the sampling is probabilistic. However, if we set quality control by retesting only the first and last blood specimens, all others have no chance of being chosen for the retesting sample, and therefore, the sample is not probabilistic. Scientific research is valid only if probabilistic samples are involved. The individuals in the sample must be independent of each other, meaning that the selection of one individual should not affect the selection of any other. The characteristics of individuals in such a sample should form distributions that are equal or similar to those in the population from which the sample is drawn. This ensures that the sample is representative and that the conclusions drawn from such a sample are applicable to the entire population it represents. (Fig. 1) (2).

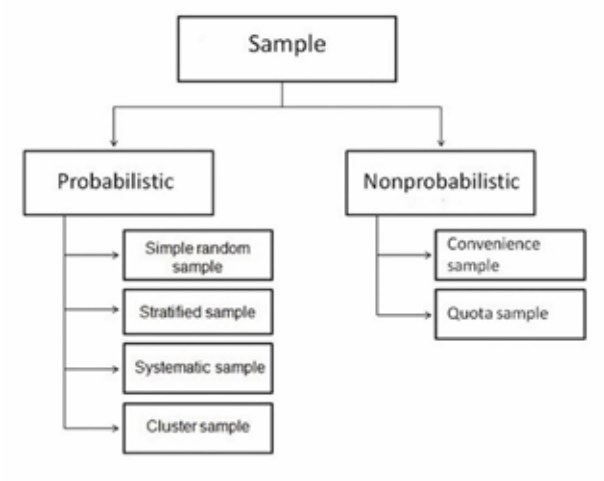

Figure 1. Sample characteristics according to a probability model.

PROBABILISTIC SAMPLES

If every subject from the population has an equal probability of being selected for the sample, i.e., the selection of one subject does not affect the selection of any other in any way, a simple random sample is formed. A simple random sample has the best population representativeness among all other sample types. The process of randomly choosing subjects is called randomization (2). For example, if a professor intends to randomly select four out of twenty students for an exam, a simple random sample could be set by blindly selecting the homework of four graded students from a pile of twenty (not by taking the first four; random selection is the point!).

A systematic sample is generated by using a systematic selection rule (for example, selecting every third, fifth, twentieth, or n-th subject from the whole) after randomly choosing the initial subject from which the selection begins. The selection according to such a system will only be effective if the population list is created without any meaningful order, ensuring that the listing does not coincide with any regularity in the population. Otherwise, the sample might not be representative (2). For example, if 100 out of 500 laboratory results are needed to be reviewed, first of all, it is necessary to determine a systematic selection rule. According to dates, a set of 1 to 5 should be chosen. One result should be randomly taken from each set (500/100=5). Otherwise, the sample would not be random.

A stratified sample is defined by subjects' randomization after the population is first split into "subpopulations" or "stratums" according to certain characteristics such as age, gender, or disease stage. Therefore, a stratified sample is a simple random sample obtained for each stratum and must be used whenever we have knowledge that the result of the study depends on the stratification indicator (e.g., hemoglobin values for women and men) (6).

If the whole population is divided into clusters and a sample is formed by randomly selecting a certain number of

clusters, a cluster sample is formed. In biomedicine, cluster samples are rarely used in clinical trials because of possible problematic representativeness. However, they are often employed in epidemiological studies where drawing a sample out of a huge population is impossible by any other method (2,6). For example, if someone wants to research the attitude of women towards prevention programs for the early detection of breast cancer in different city districts, then a cluster sample will be formed. Certain smaller districts will be accepted as ready-made clusters, and a random sample of districts will be selected where the women of all households will be surveyed. A systematic sample is selected if a large number of households is present.

NONPROBABILISTIC SAMPLES

Nonprobabilistic sample types are not based on the probability of a random selection and, therefore, might not be representative at all. Findings from nonprobabilistic samples may not be generalized to the population (2), but these samples might be suitable for studies where no statements about the population are concluded.

A convenience sample presents a selection of mostly accessible patients, relatives, volunteers, or similar subjects who are conveniently available at the time of need (6). Although not representative, it is typically used in pilot studies (for example, the first testing of a new questionnaire might be conducted among close colleagues to check the validity of the questionnaire, rather than exploring the answers).

A quota sample is commonly used for research related to market demand, e.g., gathering opinions on health care from citizens of different age groups. In this method, the researcher must determine in advance the number of people (quota) from each stratum to be interviewed. The quota sample is driven by subjective factors and, therefore, is not representative; the assumption is that only those willing to cooperate will be surveyed (6).

SAMPLING

Sampling is a process that generates a sample out of the population (2). As the sample covers only a part of the entire population, the presentation of each sample statistical parameter must be expressed with the sampling error to enable us to draw a conclusion about the population. This statistical estimate is called standard error (usually denoted by SE in scientific literature) (5,7,8). If the sample covers the entire population, there is no sampling error, and consequently, the standard error value is zero (2). Every sample statistical parameter has to be presented with its standard error; for example, the arithmetic mean has to be presented with the standard error of the mean (mean ± SEM), correlation coefficient with standard error of correlation coefficient ($r \pm SE_r$), and sample proportion with standard error of proportion ($p \pm SE_p$).

The calculation of statistical parameters and their standard errors is mathematically defined. For a simple example, it is well known that the arithmetic mean is calculated as the sum of all values divided by the number of measurements. The standard error of the mean is calculated as a ratio of the standard deviation (SD) and the square root of the number of measurements. The number of measurements (in fact, sample size, N) in the denominator indicates that the (sampling) error decreases as the sample size increases (7).

The standard error estimates the range within which the true value of a statistical parameter exists in the population. For instance, if the blood glucose in a sample of N = 30 subjects was measured as 4.8 ± 1.2 mmol/L (mean and SD), then the SEM = 0.22 and the mean glucose value in the population is 4.8 ± 0.2 mmol/L. Even more, the standard error enables us to calculate the confidence interval, an interval that will place the real value of the statistical parameter in the population with a certain probability, usually 95% or 99% (7,8,9).

Bias is an undesired occurrence of a systematic error during scientific research, and some types of biases might occur through sampling, i.e., to be a result of sampling error. For example, response bias is created by the influence of the methodology on the course of the research. A good example might be a study using a questionnaire where all participants are not equally interested in it, i.e., a survey of smokers or nonsmokers on the hazards of smoking; if applied to a sample of volunteers, it may happen that a questionnaire is filled by fewer smokers, and the sampling error may falsely increase the proportion of nonsmokers (2). Affiliation bias occurs when sample differences are interpreted only by means of known parameters when we are not aware of the possible existence of hidden population differences (i.e., better long-term health conditions in physicians at the radiology ward compared to all others in the hospital because of strict health regulations before residency; according to health status at the moment of employment, these two samples come from two distinct populations) (2).

CALCULATING SAMPLE SIZE

The most common issue related to sampling is its size – what qualifies as a big or small sample and how many subjects are needed for the overall coverage of the research.

A sample is considered large enough if it allows us to accept or reject the null hypothesis with reasonable certainty and properly assess the population parameter. Methodologically, we tend to include as many subjects from the population as possible, and the size should be adjusted to a value by which we can conclude about the population with sufficient certainty as if we were dealing with the entire population (7). Of course, despite the sample size, its representativeness must be fulfilled.

In principle, samples with fewer than 30 subjects are usually considered small, while those with more than 50 or 100 subjects are large, but the exact numerical cut-off value does NOT exist and was not published in respective literature to the best of our knowledge (2). The required sample size depends on the research type, variable variability, and the level of statistical significance (2,4). If the variability of instance is small, a small(er) sample is needed (e.g., fasting blood glucose concentration of healthy people), and if it is high, a larger sample is required (e.g., C-reactive protein concentration during acute glomerulonephritis). The same

rule applies to the precision of measurements: the higher the accuracy is needed, the larger the sample is required. The modern concept of scientific research methodology claims for exact sample size calculation, considered as a minimal required sample with distinct values of effect estimate and levels of statistical errors of hypothesis testing. The sample size might be calculated using general or quick formulae or statistical tables, graphs, software tools, and web utilities (see examples in Figs. 2 and 3), all based on general or quick formulae (7,13,14).

General formulae are complex mathematical equations, while quick formulae are designed to estimate the sample size based on relevant parameters only. Therefore, they require less data to be known and are easier to use. For example, for using mostly known Lehr's equation for sample size calculation, we have to identify levels of statistical errors α and β, and values of the sample statistical parameters, i.e., mean and standard deviation (4). Precalculated tables are published for different data and statistical hypotheses, and a good example might be a general table for sample size determination based on the size of the population presented by *The Research Advisors* on the web page (http://www.research-advisors.com/tools/SampleSize.htm) (11). Altman›s nomogram is an easy-to-use diagram appropriate for various statistical tests. For each test, we have to calculate the standardized difference (Std_D, parameter and statistics specific value; e.g., for comparison of two samples using an unpaired t-test) and join its value to the test power (i.e., 1 – β value) to read the required (minimal) sample size (7). This nomogram can also be utilized to evaluate the power of a hypothesis test for a given sample size.

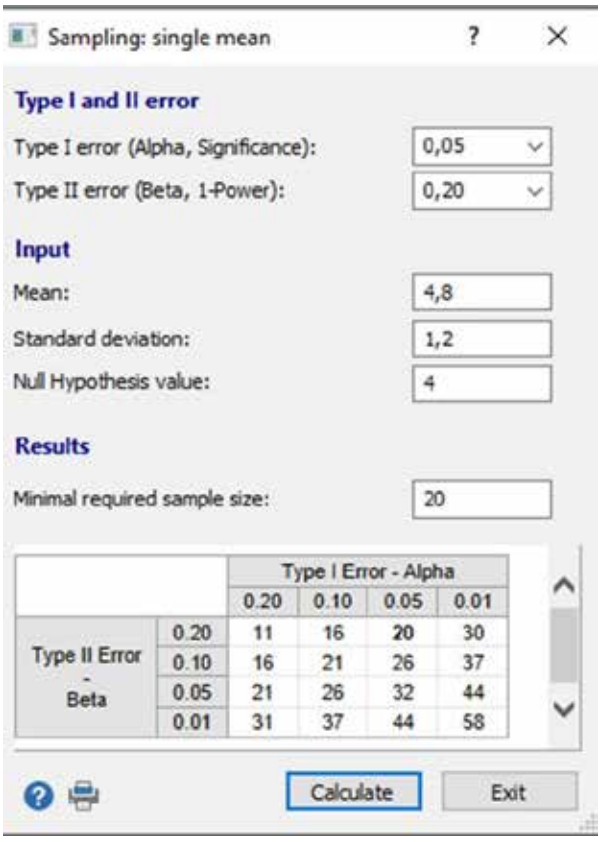

Figure 2. Sample size calculation by statistical program MedCalc.

In this example, the researcher is testing whether the blood glucose level of 4.8 ± 1.2 mmol/L is significantly different from the expected value of 4 mmol/L in the population, with 80% statistical power (b = 1 – 0.8 = 0.2) and a = 0.05. The required sample needs to be at least 20 individuals.

Computer software for statistical data analysis, such as *MedCalc Statistical Software version 16.2.1 (MedCalc Software, Ostend, Belgium;* Fig. 2), *Statistica 14.0.0.15 (TIBCO Software Inc, Chicago, United States), SPSS (IBM Corporation, United States, Armonk, New York),* and *ClinCalc* (Fig. 3), a freely available web-based tool containing frequency testing with the option of calculating sample size, are very useful in the biomedical field, Fig. 2 and 3 (13,14).

Figure 3. Sample size calculation using ClinCalc.com. when the dependent variable is a category.

Statistical power is a measure of the likelihood that the researcher will find statistical significance in the sample if the effect exists in the entire population (4). It implies the probability of correctly rejecting the false null hypothesis, thereby avoiding a type II error. This procedure is important for determining the appropriate sample size needed to detect a true effect (if it exists) with a certain level of confidence.

Power analysis is a function that involves several factors: sample size (N), effect size, significance level (α), and the power of the statistic used (1-β). The sample size includes the total number of participants involved in the research. A larger sample increases the power of the test and achieves a higher probability of detecting a true effect. Effect size seeks to establish the size of the difference or relationship the research seeks to uncover. Larger effects are easier to detect and require fewer samples, while smaller effects require a larger sample to achieve the same level of power. Furthermore, when determining the level of statistical significance, the threshold is most often set at 0.05. This level represents the probability of rejecting the null hypothesis if it is actually true (the type I error). Statistical power is usually defined as at least 80%, meaning there is an 80% or greater probability that the chosen statistical test will actually detect an existing difference in the population (4,6). Thus, the primary reason for conducting a power analysis is to determine the required sample size, thereby achieving adequate power for the research. However, a power analysis can also be performed post hoc to better understand why research may have failed to detect a significant effect. However, a pre-research power analysis can be used to ensure that the research is set up correctly from the start. Striking a balance between sample size, effect size, significance level, and statistical power is key in biomedical research. A properly planned power analysis can influence the reliability and validity of research results and help researchers make the right conclusions (2,4).

CONCLUSION

The sample is a subset of the population obtained through the process of sampling used in specific scientific research. Conclusions drawn from a sample must be generalized to the entire population from which the sample was selected. Research must be carefully planned, including many aspects such as sample size calculation and power analysis (12). Sampling creates systematic errors. To keep it minimal, the sample must be representative, and all possible biases should be avoided. Good representativeness in the sample is achieved by using random selection and proper sample size. The sample size is the only factor that researchers can control. It directly affects the statistical power of the test in each study. Thus, the larger the sample (assuming it is representative), the greater the statistical power.

ACKNOWLEDGMENTS

I would like to thank the late Full Professor Mladen Petrovečki, Ph.D., for his support in writing of this paper.

REFERENCES

1. Khaled Fahim N, Negida A. Sample Size Calculation Guide - Part 1: How to Calculate the Sample Size Based on the Prevalence Rate. Adv J Emerg Med. 2018 Jul 31;2(4).
2. Petrovečki M. Sample and population. In: Marušić M, ed. Principles of Research in Medicine. *6th* Zagreb: Medicinska naklada; 2019. p. 61-72.
3. Yildirim K, Uzkeser H, Keles M, Karatay S, Kiziltunc A, Dursun Kaya M. Relationship between serum interleukin−1β levels and acute phase response proteins in patients with familial Mediterranean fever. *Biochem Med 2012;22(1):109-13.*
4. *McHugh ML. Power analysis in research. Biochem Med 2008;18(3):263-74.*
5. Zhang X, Hartmann P. How to calculate sample size in animal and human studies. Front Med (Lausanne). 2023 Aug 17;10:1215927
6. *Petz B. [Osnove statističke metode za nematematičare]. 5th ed. Jastrebarsko: Naklada Slap; 2012.*
7. Althubaiti A. Sample size determination: A practical guide for health researchers. J GenFam Med. 2023;24:72–7
8. *Das S, Mitra K, Mandal M. Sample size calculation: Basic principles. Indian J Anaesth. 2016 Sep;60(9):652-656.*
9. *Šimundić AM. Confidence interval. Biochem Med 2008;18(2):154-61.*
10. Ilakovac V. Statistical hypothesis testing and some pitfalls. Biochem Med 2008;19(1):10-6.
11. *Sample size Table. Available at:* http://www.research-advisors.com/tools/SampleSize.htm. Accessed May 9, 2024.
12. *Motulsky H. Choosing an Appropriate Sample Size. In: Motulsky H, ed. Intuitive Biostatistics. New York-Oxford: Oxford University Press; 1995. p. 195-204.*
13. *MedClac.org Sample size calculation. Available at: https://www.medcalc.org/index.php. Accessed May 29, 2024.*
14. *ClinCalc.com. Sample Size Calculator. Available at: https://clincalc.com/stats/samplesize.aspx. Accessed May 9, 2024.*