Tools for data analysis

Santiago Perez-Lloret MD PhD [1,2,3], Alejandro Enet PhD [1], Gabriela Gonzalez-Alemán PsyD, PhD [4].

[1] Observatorio de Salud, Pontificia Universidad Católica Argentina, Alicia Moreau de Justo 1300, Buenos Aires, Argentina
[2] Consejo de Investigaciones Científicas y Técnicas (CONICET), Argentina.
[3] Departamento de Fisiología, Facultad de Medicina, Universidad de Buenos Aires (UBA), Buenos Aires, Argentina.
[4] Facultad de Psicología y Psicopedagogía, Pontificia Universidad Católica Argentina, Alicia Moreau de Justo 1300, Buenos Aires, Argentina

Running title: Tools for data analysis

Word count: Text= 2300, Figures= 2, Tables= 1.

Corresponding author:
Santiago Perez-Lloret MD, PhD
Observatorio de Salud, Pontificia Universidad Católica Argentina
Av. Alicia Moreau de Justo 1300, Buenos Aires, Argentina
Telephone / Fax: +54 11 810 2200 822
e-mail: santiagopl@conicet.gov.ar

What are statistics Good for in human research studies?

Studies conducted on human beings may have different objectives and designs, but they all share some common principles.[1] We outline these principles as a cycle, shown in Figure 1.

The first step is to obtain a sample from a population. A population is a group of human beings sharing one or more characteristics. In medical research, researchers usually define populations following a disease or a condition. Obtaining the sample is called “sampling”.[2]

Researchers will then discuss the study with the potential participants. They will be part of the study sample if they accept to participate and fulfill all inclusion and exclusion criteria. Investigators will perform a series of procedures and assessments and may apply an intervention to the sample of participants. For example, a treatment may be used, and its effects on Parkinson’s Disease motor symptoms may be recorded. Notably, study results only represent the effects of the intervention on the sample of participants. However, researchers are generally interested in “extrapolating” these results to the target population. The “statistical inference” procedure allows for performing such extrapolations.[3]

Statistics is the science of collecting, analyzing, and describing data to conclude a particular phenomenon based on a relatively limited sample material.[3] It employs mathematical and probabilistic tools to develop methods and models for data analysis. Statistics is a highly interdisciplinary field; research in statistics finds applicability in virtually all scientific fields. Statisticians will intervene mainly in three crucial steps of the human research study cycle: sampling, results description, and inference. We will briefly review these topics in the following paragraphs.

Statisticians should be part of every human research team to handle these tasks [4].

Sampling

Sampling in human research involves selecting a part of the population to obtain the necessary data for analysis.[2] Remember that a population is a group of beings with one or more characteristics. A population includes existing beings as well as those that exist and will exist in the future. In this sense, populations are infinite, and thus, sampling is necessary to explore their characteristics. The sample is the specific group of individuals from whom you will collect data.

The process of sampling starts with the definition of the population. Usually, only a subset of the population of interest is targeted. Subjects may be excluded from the "target" population because they have characteristics that would produce bias or safety issues. These are typically defined in the study protocol. Another common reason for exclusion, particularly in studies employing non-random samples, is that the researchers cannot reach subjects or are unwilling to participate.

There are two main sampling methods. Probability sampling involves random selection, allowing solid statistical inferences about the whole group. Conversely, non-probability sampling involves non-random selection based on convenience or other criteria. It will enable accessible data collection but may introduce bias. It is the method of sampling most commonly used in human research studies.

Another crucial aspect of sampling is determining sample size. Studies with small samples will lack the "statistical power" to achieve their goals. On the other end of the spectrum, too many subjects will increase study costs and burden without providing further benefits in terms of representativity or validity. The number of individuals in the sample depends on factors like the intended effect size (i.e., the force of the association between two variables), variability, and research design. More information on this topic and simple formulas can be found in the following review articles.[5,6] Nowadays, the software performs sample size calculations. G*Power is a freeware, validated calculator.[7,8]

Researchers should clearly explain their sampling method in the methodology section of research papers, signal any potential bias arising from sampling, and explain the assumptions used to calculate the sample size.

Descriptive statistics

Once data collection is finished and the study database has been audited, corrected, and closed, it's time to describe the results! The objective is to provide an appropriate, concise, and clear summary of the findings. Successively, tables and graphs should show the sample characteristics and the primary and secondary study outcomes.[9] John Tukey (1915-2000), an outstanding mathematician and statistician, once said that a simple graph may provide more information to the data analyst's mind than any other device. His book Exploratory Data Analysis, written in the pre-intensive computing era, is an exciting reading.[10] Tables help convey more extensive information and can be very effective if kept simple.

One can organize data in a meaningful way that allows the researcher to identify patterns, trends, and relationships within the data, making it easier to explore and hypothesize for further analysis.

The importance of descriptive statistics cannot be overemphasized. According to John Tukey, once upon a time, statisticians only explored.[10] This changed during the XX century when confirmatory "inferential" analysis techniques were developed.[4] However, nothing that hasn't been previously visualized can be confirmed. Exploratory and confirmatory analyses should proceed side-by-side. We shall discuss inferential statistics in the next section.

Inferential statistics

Inferential statistics involve making predictions or drawing conclusions about a population based on data collected from a sample.[11] While descriptive statistics summarize characteristics of a data set, inferential statistics allow researchers to make inferences beyond the sample. Let's see how it works.

Lin and colleagues conducted a randomized, double-blind, placebo-controlled study on the effects of lovastatin for slowing motor symptoms progression in patients with early-stage PD.[12] Forty patients received a placebo and 37 lovastatin 80 mg/day for 48 weeks. MDS-UPDRS motor score in the lovastatin group changed by −3.18 ± 5.50 (Mean ± Standard Deviation) versus -0.50 ± 6.11 in the placebo group. The publication includes a nice graph showing the effects of placebo and lovastatin through the 48 weeks. This descriptive analysis shows a greater reduction in motor scores in the patients treated with lovastatin than in those on placebo. Can this result be confirmed? We need to extrapolate the results to the population to find this out.

Inferential statistics involve estimating population parameters (e.g., mean, proportion) based on sample statistics or testing statistical hypotheses.[11] Estimating population parameters is usually done by calculating confidence intervals.[11]

A confidence interval is a range of values that provides an unbiased estimate of an unknown parameter of a population with a certain level of confidence. It represents the uncertainty around a point estimate, such as a mean or proportion. It provides a range of values within which one can expect the estimate to fall if the experiment is repeated. An important concept is the confidence Level. If we construct a 95% confidence interval, we are confident that 95 out of 100 times, the estimate will fall between the upper and lower values specified by the interval. A 95% confidence

level is usually used in human research. We can calculate confidence intervals for various statistical estimates, including proportions, means, differences between population means or proportions, and estimates of variation among groups.

We want to calculate the 95% confidence interval for lovastatin's effect in the study by Lin and colleagues. The difference between treatments is -2.68. The 95% confidence interval for this point estimate is -5.33 to 0.00. This indicates that there is a considerable possibility that the true difference between lovastatin and placebo in the population is 0, which means that the drug might not have any effect.

Statistical hypotheses testing is a complementary technique for statistical inference. Researchers can make inferences about the population by testing specific hypotheses arising from scientific theories.[4] Usually, two hypotheses will be contrasted. The "null hypothesis (H_0)" predicts that the characteristics of the variables are mostly the same in all situations considered (e.g., there is no difference in the mean value of two samples of cases). Conversely the "alternate hypothesis (H_1)" predicts differences in these characteristics (e.g., there is a difference in the mean of the two sample of cases). H0 and H1 might be inverted in some cases, such as in non-inferiority clinical trials. For example, let's say we are testing whether men are, on average, taller than women. H_0 would be "Men are, on average, not taller than women," and H1 would be "Men are, on average, taller than women." We would then measure a sample of men and women and perform the statistical test. The description of the statistical testing procedures is beyond the scope of this manuscript. However, interested readers might consult the Fundamental of Biostatistics book by B. Rosner.[3] It suffices to say that based on the results of the calculations, i.e., the p-value (i.e., the probability of the null hypothesis being true), the researcher can determine if there's enough evidence to reject the null hypothesis and accept the alternate one. Hypothesis testing helps researchers make informed decisions by assessing the significance of differences, relationships, and patterns in data.

In the study from Lin and colleagues, H_0 would be "motor symptom progression with lovastatin = placebo," whereas H_1 would be "motor symptom progression with lovastatin < placebo." The author reports a p-value for this difference of 0.14. Conventionally, H_0 is rejected when the p-value is < 0.05 (called the "α critical value"). Therefore, unfortunately, the hypothesis stating that "motor symptom progression with lovastatin ≥ placebo" cannot be rejected. A critical issue with

“negative” or “non-significant” results is the lack of statistical power, which can lead to a “false-negative” result, called “β error”. Including a large enough sample is the only protection against this error.[13]

Results from confidence intervals are usually coherent with those of statistical hypotheses testing. It is recommended that both are reported, as confidence intervals can go beyond the simple p-value by highlighting the uncertainty of the results.[14]

Surviving the Titanic

On April 15, 1912, during her maiden voyage, the widely considered “unsinkable” RMS Titanic sank after colliding with an iceberg. Unfortunately, there weren’t enough lifeboats for everyone onboard, resulting in the death of 1,502 out of 2,224 passengers and crew. A dataset containing the characteristics and faith of 891 persons on board is available publicly (https://www.kaggle.com/datasets/yasserh/titanic-dataset). We will use the subset of cases with complete data from this dataset to propose a framework for scientific data analysis. As shown in Table 1, 424 out of the 714 cases died (59.4%). When navigating data tables, the first step is identifying the research outcomes and their determinants. In observational studies like this one, the outcome will typically be shown in columns, and the risk factors will be displayed in the rows. In the case of clinical trials, outcomes are frequently shown in the rows and the intervention in the columns. All statistical analyses in this example were performed by R v4.2.0 (The R Foundation for Statistical Computing, Vienna, Austria).

The outcome of this study is survival. The second step is visually exploring data. Survivors were more frequently females, somewhat younger, traveled in larger kindreds, traveled in First class, for which they paid a higher fare, and embarked more frequently in Cherbourg, France. The result of the statistical hypothesis test for each variable is shown in the fourth column, “Unadjusted p-value.” It is important to check whether statistical hypothesis tests were adequately employed.

We have previously discussed the “false-negative error”. The opposite can also happen, that is, rejecting H_0 and accepting H_1 when H_0 is true. This is called the “false-positive α error”. The lower the critical α level selected, the fewer the chances of the false-positive α error. When the critical α value is arbitrarily fixed at 0.05, the false-positive error risk is 5%. This discussion shows that the chance of finding a

false-positive risk in two comparisons is 5% multiplied by 2 = 10%. Table 1 shows six comparisons, which results in an “experiment-wise” false-positive error of 6 times 5%, thus 30%. In other words, our analysis has a 30% chance of finding at least one false-positive result, which is unacceptably high. Several techniques are used to “adjust” for multiple comparisons.[15] The interested reader might want to read the excellent article from John Ludbrook.[16] The fifth column (“Adjusted p-value”) shows the p-value adjusted for multiple comparisons using the Holm method.

The last step is to identify the independent and significant predictors of survival in the sample. In other words, are there any confounding relationships among the variables? Confounding is when the observed relationship between an independent variable (exposure) and a dependent variable (outcome) is distorted or biased by the presence of one or more variables, known as confounders. For example, is the relationship between age and survival confounded by sex? The inverse may sound more familiar, but are age and sex independent predictors of survival? Researchers use multivariate statistical techniques to answer these questions. In this case, a logistic regression analysis showed that the independent and significant predictors of survival were female gender, younger age, and class. Fare was not included in this analysis as it was hypothesized to relate closely to class.

Using Machine Learning algorithms in human research studies

Machine learning (ML) is a subfield of artificial intelligence that uses algorithms trained on data sets to create self-learning models.[17] These models can automatically predict outcomes, classify information, and perform tasks that would otherwise require human intervention. The applications of ML in health are evolving fast.[17] However, the usefulness of these algorithms for human research is less clear. Explainability refers to the ability to understand and interpret the ML model's outcomes. Human research is all about explainability. However, most ML models work in a “black box”, difficult to interpret fashion.[18] Despite this, ML models may still have a place in human research. For example, we fitted a decision tree predicting survival (Figure 2). Results go beyond the logistic regression, as it explores the interactions between variables. By analyzing Figure 2, we learn that gender is the most important predictor of survival. Furthermore, there is a difference in the set of additional predictors in men and women. Finally, third-class passengers who paid a lower fare had better survival rates. In summary, ML methods can provide valuable

insights into the risk factors of the outcome, besides the widely used logistic regression models.

Author's Roles

1. Research project: A. Conception, B. Organization, C. Execution;
2. Statistical Analysis: A. Design, B. Execution, C. Review and Critique;
3. Manuscript Preparation: A. Writing of the first draft, B. Review and Critique;

SPLL: 2A, 2B, 3A
AE: 2C, 3B
GGA: 3B

Disclosures:

Authors have no conflict of interests for this work.

Financial Disclosures for the previous 12 months:

SPLL have received honoraria from the IPMDS and grants from the Agencia de Promoción Científica y Técnica (Argentina).
AE and GGA declare that there are no additional disclosures to report.

Ethical Compliance Statement:

This is a review article that did not involve patients. IRB approval is not needed and informed patient consent not relevant.
We confirm that we have read the Journal's position on issues involved in ethical publication and affirm that this work is consistent with those guidelines.

References

1. Elwood M, editors. Critical appraisal of epidemiological studies and clinical trials, Oxford, Oxford University Press, 2017;
2. Flight L, Julious SA. Practical guide to sample size calculations: an introduction. Pharm Stat 2016;15:68-74.
3. Rosner B, editors. Fundamentals of biostatistics, Boston, MA, Cengage Learning, 2016;
4. Ludbrook J. Statistics in physiology and pharmacology: a slow and erratic learning curve. Clin Exp Pharmacol Physiol 2001;28:488-492.
5. Julious SA. Sample sizes for clinical trials with normal data. Stat Med 2004;23:1921-1986.
6. Julious SA, Campbell MJ. Tutorial in biostatistics: sample sizes for parallel group clinical trials with binary data. Stat Med 2012;31:2904-2936.
7. Faul F, Erdfelder E, Buchner A, Lang AG. Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses. Behav Res Methods 2009;41:1149-1160.
8. Faul F, Erdfelder E, Lang AG, Buchner A. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Methods 2007;39:175-191.
9. Larson MG. Descriptive statistics and graphical displays. Circulation 2006;114:76-81.
10. Tukey JW, editors. Exploratory data analysis, Reading, Mass., Addison-Wesley Pub. Co., 1977;
11. Sullivan LM. Estimation from samples. Circulation 2006;114:445-449.
12. Lin CH, Chang CH, Tai CH, et al. A Double-Blind, Randomized, Controlled Trial of Lovastatin in Early-Stage Parkinson's Disease. Mov Disord 2021;36:1229-1237.
13. Altman DG, Bland JM. Absence of evidence is not evidence of absence. BMJ 1995;311:485.
14. Ho J, Tumkaya T, Aryal S, Choi H, Claridge-Chang A. Moving beyond P values: data analysis with estimation graphics. Nat Methods 2019;16:565-566.
15. Schulz KF, Grimes DA. Multiplicity in randomised trials I: endpoints and treatments. Lancet 2005;365:1591-1595.
16. Ludbrook J. Multiple comparison procedures updated. Clin Exp Pharmacol Physiol 1998;25:1032-1037.
17. Verma VK, Verma S. Machine learning applications in healthcare sector: An overview. Materials Today: Proceedings 2022;57:2144-2147.
18. Hassija V, Chamola V, Mahapatra A, et al. Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence. Cognitive Computation 2024;16:45-74.

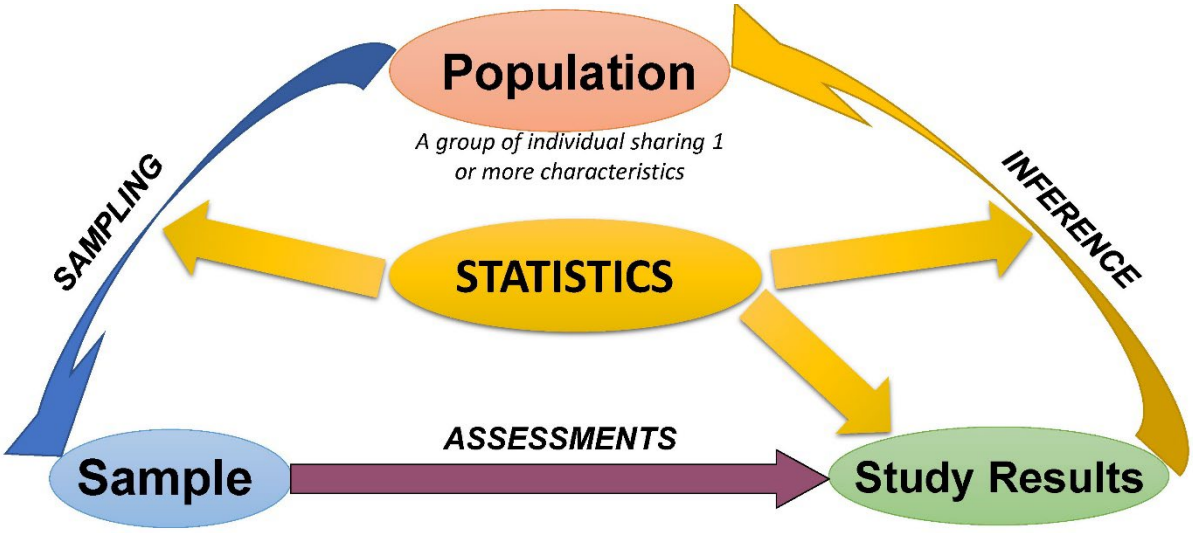

Figure 1. Cycle of a human research study.

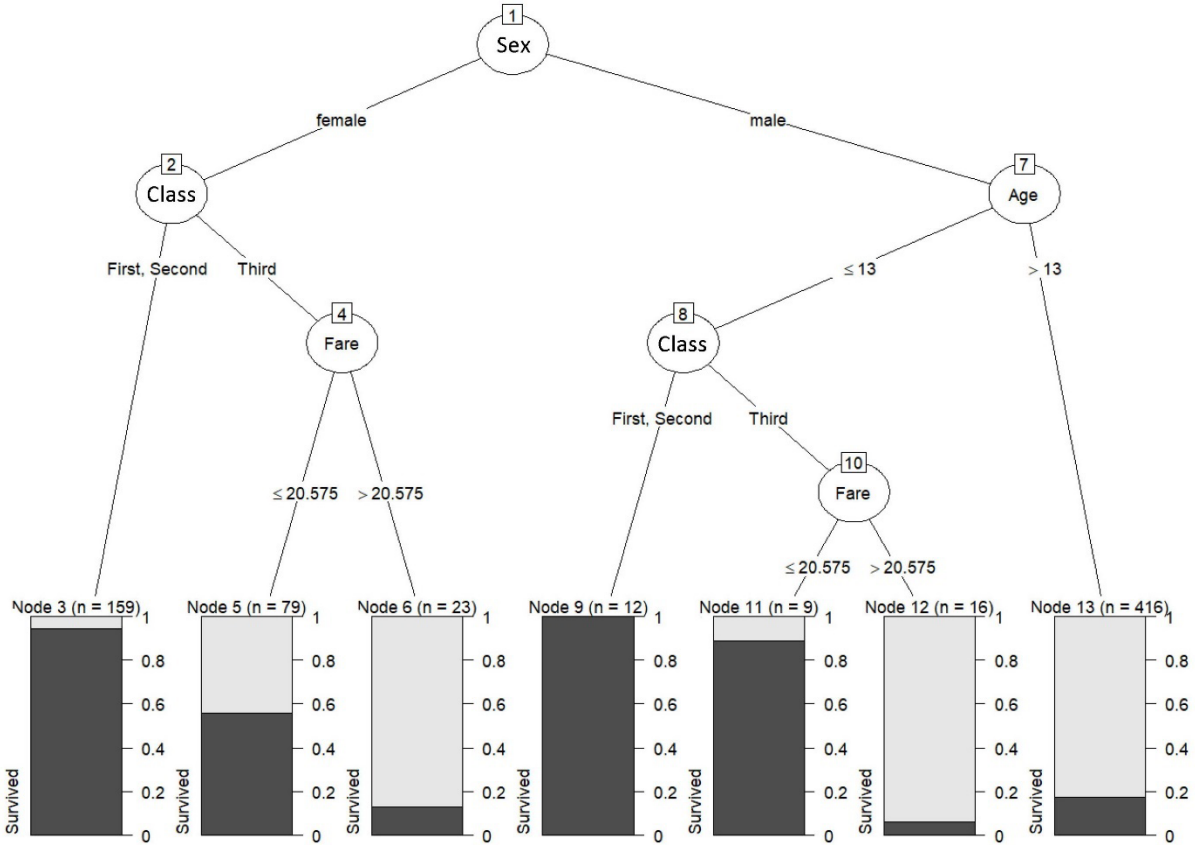

Figure 2. A decision tree to predict the survival of Titanic passengers.

Table 1. Risk factors for death among Titanic passengers.

	Died (n=424)	Survived (n=290)	Unadjusted p-value	Adjusted p-value	Logistic regression OR (95% CI)
Sex					
Female	64 (15.1%)	197 (67.9%)	<0.001	<0.001	1
Male	360 (84.9%)	93 (32.1%)			0.07 (0.05-0.11)*
Age	30.6±14.2	28.3±15.0	0.039	0.039	0.96 (0.95-0.98)*
Family size	0.34±0.87	0.53±0.81	0.011	0.022	0.85 (0.66-1.07)
Class					
First	64 (15.1%)	122 (42.1%)	<0.001	<0.001	1
Second	90 (21.2%)	83 (28.6%)			0.32 (0.17-0.61)*
Third	270 (63.7%)	85 (29.3%)			0.09 (0.04-0.18)*
Fare paid	23.0±31.4	51.8±70.5	<0.001	<0.001	Not included
Port of embarkation					
Cherbourg	51 (12.0%)	79 (27.2%)	<0.001	<0.001	1
Queenstown	20 (4.7%)	8 (2.8%)			0.43 (0.13-1.30)
Southampton	353 (83.3%)	201 (69.3%)			0.62 (0.36-1.05)

Bivariate test p-values were adjusted by employing the Holms procedure. * p<0.001 (Wald test).