REVISTA CIE

Revista de Investigación Educativa | Vol. 1, No. 2
ISSN: 3061-8657 | Julio-Diciembre 2025

STATISTICS

Analysis and interpretation of data in medical research

Dr. Gerardo García Maldonado
Doctoral Candidate in Health Sciences
Full-Time Professor SNII-I
Universidad Autónoma de Tamaulipas
Facultad de Medicina de Tampico "Dr. Alberto Romo Caballero"
gmaldonado@docentes.uat.edu.mx

Dr. Wilberto Sánchez Márquez
PhD in Educational Development
Full-Time Professor SNII-I
Universidad Autónoma de Tamaulipas
Facultad de Medicina de Tampico "Dr. Alberto Romo Caballero"

INTRODUCTION

The perception of statistics among students at the Faculty of Medicine of Tampico "Dr. Alberto Romo Caballero" varies considerably, but some common aspects stand out. For example, many of them recognize the importance of this discipline in medicine in general, and in scientific research in particular, considering the wide range of medical information that is now available on electronic platforms, much of which is expressed in mathematical terms, probabilities, estimates, or statistical significance. Of course, I am referring to scientific information. Today, a significant theoretical and practical foundation in this field is required for adequate critical reading of scientific literature. It is recognized, therefore, that it is also a fundamental tool for evidence-based clinical decision-making.

Secondly, although students are aware of its importance, most find it difficult to understand, which generates anxiety when facing this subject, as it is part of the academic curriculum in medical school. Without a doubt, the terminology used and abstract concepts, as well as the equations, calculations, formulas, and algorithms this implies can be intimidating. There are, of course, other contributing factors. On one hand, the teaching of this subject by professionals who are not physicians, which, while it might presuppose that as experts they would be better suited, in reality does not work as the approach, whether intended or not, is different. The experience of being a physician makes a difference. But on the other hand, and perhaps most significantly, is the lack of interested medical faculty with competencies and experience in this field.

Nowadays, quantification is essential for data to be applied appropriately and conveniently. Experts agree that a clinician or medical researcher who is not able to numerically express results or conclusions will not provide reliable information. In other words, statistics as a quantitative analysis tool is, in this case, at the service of medicine.

This statement may seem exaggerated, but the reality is that the medical information we review within our fields of competence today bears the stamp of evidence-based medicine, where the critical analysis of the literature focuses on statistical and methodological aspects for diagnostic, prognostic, pharmacological, non-pharmacological and preventive decision making. David Sackett and Gordon Guyatt, both academics at McMaster University in Ontario, Canada, noticed as early as the 1970s how inaccessible it was for clinicians to read the enormous amount of information being generated; it was necessary to establish practical, critical and relatively agile strategies for scientific reading.

To this end, they deemed it appropriate to train professionals and provide them with knowledge in research methodology and biostatistics; therefore, in the 1990s David Sackett together with clinical epidemiologists from that university, began a series of guide-articles in the U.S. journal JAMA about the critical appraisal of medical literature. It goes without saying how successful their work was. It is worth having the full series of those publications. We recommend searching for and reading them.

Over time the need to practice agile and dynamic reading not only persists, but has increased exponentially. The evidence is overwhelming: it is imperative to know how to read a scientific paper in order to evaluate the best options for the benefit of patients. Of course this is for medical purposes; but what if we do not only want to be receivers of knowledge, but also aspire to be generators of new knowledge? In that case we will equally need statistical and methodological competencies.

As mentioned in the first booklet, some procedures will be reviewed to master the appropriate use of statistical tools, as well as to know and understand when and how to use a given method that allows interpretation of information and making the best decisions. In this second issue we present material with three objectives:

1. First, to expose a taxonomy of statistics that is simple and practical to understand, adhering to the principle of parsimony.
2. Second, to conceptualize and contextualize each of the elements that make up the taxonomy. A more integrative and detailed breakdown will be provided in subsequent booklets.
3. Finally, to present a glossary of symbols that is commonly found in mathematics in general and statistics in particular.

TAXONOMY IN STATISTICS

A taxonomy is an organized classification system that groups and categorizes elements based on common characteristics. It is often considered synonymous with classification, but for the purposes of this project the term taxonomy will be used. The term comes from ancient Greek: (taxis, "ordering") and (nomos, "norm" or "rule"). It is a procedure applied in science in general as well as in other fields of knowledge. The ultimate purpose, wherever required, is to organize and understand information. Taking these considerations into account, statistics is no exception.

For a taxonomy to be effective it must be clear, simple, relevant and compatible with norms or standards, because that way it will ensure efficient use of information. Reviewing the literature and contrary to what might be thought, a single taxonomy in statistics does not exist.

On the contrary, there are diverse proposals and often statistical aspects are mixed with topics related to research methodology. Of course these elements are complementary in scientific research, so it is really difficult to separate them. But the idea on this occasion is to try.

It is not the primary objective of this booklet to propose new classifications or something atypical; the suggestion is simply an organization that addresses exclusively the statistical theme. The intention in this document was to adhere to the principle of parsimony, which states that under equal conditions the simplest explanation is usually the best and most convenient.

Taking into account the intrinsic characteristics of the elements indicated in this taxonomy, an attempt was made for it to be exhaustive and mutually exclusive, that is, to leave no element outside the formed groups and that those elements belonged only to one group. Based on the above, the following is presented: STATISTICS (main branches)

- DESCRIPTIVE
- INFERENTIAL
- PARAMETRIC
- NON-PARAMETRIC

DESCRIPTIVE STATISTICS. CONCEPTS AND GENERALITIES

Descriptive statistics is a key piece of data analysis. As the name implies, its objective is to describe and summarize the characteristics of the information provided so that it is easy to understand. This strategy does not aim to make inferences or draw aggregated conclusions.

This type of analysis involves the computation and presentation of various statistical measures, such as the mean, median, mode, standard deviation, variance and percentiles, which provide valuable information about central tendencies, dispersion and the overall distribution of data. Another main objective of descriptive statistics is to communicate as much information as possible in the simplest and most efficient way.

By presenting these summarized measures, readers can obtain a clear picture of the data, including typical values, the degree of variability, and any possible outliers or skewness in a sample or a population.

Its use extends to several fields, including data science, where it plays a crucial role in the initial exploration and understanding of data. In the context of medical research, it is used to summarize participant characteristics such as age, race, gender, social stratum and disease prevalence, to name a few examples.
By understanding this information, informed decisions can be made to perform further analyses and hypothesis testing with more advanced statistical methods, which will lead to more reliable conclusions.

Descriptive statistics has the advantage of representing obtained data in bar charts, histograms, scatter plots or line charts, among other methods, which further facilitates understanding of the information.

We can summarize, then, that the objectives of this statistical strategy are:

- Collect data
- Organize information
- Present and tabulate data
- Analyze results

INFERENTIAL STATISTICS. CONCEPTS AND GENERALITIES

Inferential statistics is a branch of statistics that allows generalizations to be made and informed decisions to be taken based on representative samples from a particular population. The condition is that the study samples are obtained from the target population through calculation, and that the study units have been selected randomly.

Unlike descriptive statistics, which focuses on ordering and describing collected data, inferential statistics uses these data as a basis for making predictions or inferences, that is, generalizations. One of the fundamental pillars of inferential statistics is the concept of probability, which provides a framework for quantifying uncertainty; this is essential because no datum is necessarily definitive — there will always be circumstances that generate errors.

Topics such as confidence intervals, parameter estimation and hypothesis testing are key elements that make it possible to quantify results and make predictions with greater certainty for entire populations from selected samples. In any case, when the sample is selected through randomization, any conclusion will be subject to a margin of error, which must be considered beforehand.

Hypothesis testing is a central procedure in inferential statistics used to determine whether there is sufficient evidence in a sample of data to infer that a certain condition is true for the whole population. This process involves formulating a null hypothesis and an alternative (research) hypothesis, and then using statistical methods to determine which of these hypotheses is more consistent with the collected data.

Confidence intervals, on the other hand, should ideally accompany the quantification of statistical parameters (measures of central tendency and dispersion) and population parameters (rates, indices, prevalence, incidence, odds ratios, relative risk, etc.), as they assess the margin of error associated with the calculation.

A 95% confidence interval, for example, suggests that if sampling were repeated many times, in 95% of the calculations the intervals obtained would contain the true population parameter. This methodology is crucial since, in the real world, because it is impracticable to approach complete data for an entire population, obtaining study samples and making inferences is more the rule than the exception. Inferential statistics is also applied in medicine — for example, it is used to evaluate the effectiveness of new treatments or medications.

It is crucial to bear in mind that inferential statistics is not free of limitations: inadequate sampling can lead to bias and, consequently, to incorrect inferences. Moreover, incorrect interpretation of statistical results can lead to erroneous conclusions and poor clinical decisions.

In summary, inferential statistics is a powerful tool that allows limited data to be turned into applicable knowledge to understand and improve the world around us. Through hypothesis formulation, probability calculation and constructing confidence intervals, it provides a robust framework for decision-making under uncertainty. However, its effective application requires a deep understanding of its methods and limitations to avoid errors and maximize its potential.

PARAMETRIC STATISTICS. CONCEPTS AND GENERALITIES

Parametric statistics is a fundamental branch of inferential statistics that focuses on the analysis of data where the data must satisfy certain rules or assumptions (as they are called) for the results to be reliable and valid. This is basic because the use of parametric statistical tools guarantees a more robust and precise analysis of the data under study. In other words, statistical tests called "parametric" are based on a set of parameters that must be met without question.

One of the central aspects of parametric statistics is the use of well-known distributions, such as the normal, binomial and Poisson distributions, which can be represented using tables, graphs or mathematical functions. In the case of the normal distribution, which is perhaps the best known, we must understand the concept as a theoretical model in which the values of a variable or set of variables under study are symmetrically distributed around a central value.

The distribution is characterized by a bell shape; it is described mathematically and depends on two parameters: the mean and the standard deviation. These distributions allow calculations to be simplified and conclusions to be obtained more quickly, provided that the data adequately fit the assumptions of the model.

The normal distribution is the basis for numerous parametric methods due to its prevalence in nature and its relationship with the central limit theorem. This theorem states that if many samples are taken from a population — for example, the height of people — and you calculate the mean of each sample, as the number of samples increases the distribution of those sample means will approach a normal curve, regardless of how the original data themselves were distributed. In other words, the larger the sample size, the closer the distribution of the means will approximate a normal distribution.

The use of parametric methods comes with several advantages: first, when the rules or assumptions for their use are met, these methods tend to be very efficient, providing precise and powerful estimates with smaller sample sizes compared to their nonparametric counterparts. Furthermore, because of the mathematical structure of parametric models, extensive analyses are facilitated that allow a deeper and more rigorous understanding of the data. However, parametric statistics is not without limitations: the main drawback lies in the dependence on assumptions about the distribution of the data. If these assumptions are not met, results may be incorrect; therefore, it is crucial to verify the distribution prior to application.

Thus, in situations where it cannot be guaranteed that the data follow a specific distribution, nonparametric methods may be more appropriate, albeit at the cost of lower efficiency. Among the most common techniques

used in parametric statistics are hypothesis tests, confidence interval estimation and analytic regression models, among others. Hypothesis testing, where statistics such as Student's t and analysis of variance (ANOVA) can be used, allow the determination of the significance of results based on specific parameters such as the population mean or variance.

Confidence intervals, in turn, provide a plausible range for unknown parameters, offering crucial information about their precision and reliability. Regression models are a powerful tool to model, among other things, the relationship between variables and to make predictions. In conclusion, parametric statistics constitute an essential tool in data analysis, equipped with efficient and powerful methods to investigate and understand natural and scientific phenomena. Although the greatest emphasis is placed on the normality of the distribution, there are other considerations that must be taken into account.

Common assumptions for many parametric methods include:

- Normality in the distribution of the data
- Homogeneity of variances (also called homoscedasticity)
- Independence among observations
- Linearity in the case of regression models
- Variables should be measured on an interval or ratio scale.

NON-PARAMETRIC STATISTICS. CONCEPTS AND GENERALITIES

Non-parametric statistics has become an invaluable tool in data analysis, especially in situations where the traditional assumptions outlined above cannot be confirmed, making it more versatile and applicable in a variety of real-world scenarios. It is important to understand that what distinguishes non-parametric statistics from its parametric counterpart is that it does not rely on specific parameters (such as the mean and standard deviation) nor does it assume a normal distribution of the data. This matters because in practice the assumptions are not always met.

Major non-parametric techniques include rank tests such as the Wilcoxon signed-rank test, the Mann–Whitney U test, and the Kruskal–Wallis H test. These tests compare medians or ranks instead of means, which makes them more robust in cases of skewed data or outliers. For example, when comparing the effectiveness of two medical treatments, data may not be normally distributed due to variability in patients' responses. In such cases a non-parametric test can provide more reliable results. The flexibility of non-parametric statistics also extends to its ability to handle different types of data, including ordinal and categorical data.

Often in social and psychological research, data are collected on ordinal scales, for example those that measure patient satisfaction on a scale from "very dissatisfied" to "very satisfied". Non-parametric methods can analyze these data without needing to convert them into a numerical scale, thereby preserving the integrity of the original measurement. A critical aspect of non-parametric techniques is their interpretative simplicity: by not depending on complex assumptions about the distribution, the results can be easier to interpret and communicate to a non-technical audience. This is particularly beneficial in interdisciplinary contexts where clarity and accessibility of results are essential.

Despite their many advantages, non-parametric methods often require larger samples to achieve acceptable statistical power; additionally, some non-parametric methods can pose challenges for very large datasets. In conclusion, non-parametric statistics offer a powerful and flexible alternative to traditional parametric methods. Their ability to analyze data without strict distributional assumptions makes them especially useful in disciplines with varied and hard-to-model data. Although they may be less efficient in terms of sample size, their applicability and ease of interpretation often make them the preferred choice in real-world studies where ideal conditions are rarely found.

GLOSSARY OF SYMBOLS

Sample standard deviation	S
Population standard deviation	σ
Sample variance	S^2
Population variance	σ^2
Population mean	μ
Sample mean	$\bar{x}$
Median	Md
Mode	Mo
Absolute frequency	fai
Cumulative absolute frequency	Fai
Relative frequency	fri
Cumulative relative frequency	Fri
Frequency of a score or category	*f*
Degrees of freedom	gl
Alternative hypothesis	H1
Null hypothesis	Ho
Significance level	α
Population size	N
Sample size	n
Summation operator	∑
Probability of committing Type II error	β
Probability of committing Type I error	α
Probability (also sometimes represents significance level)	ϱ
Proportion	P
Analysis of variance (ANOVA)	F
Z-score	z
Chi-squared	X^2
Spearman correlation coefficient	rho
Pearson correlation coefficient	r
Kruskal–Wallis test	H
Wilcoxon signed-rank test	T
Cochran's Q test	Q
Student's t-test	t
Mann–Whitney U test	U
Proportion	P
Cramér's V coefficient	C
Kappa statistic	K
i-th value of a variable	Xi
Maximum value of a variable	Xmax
Minimum value of a variable	Xmin
Variable to compute	x

Upper limit of an interval	Ls
Lower limit of an interval	Li
Width (amplitude) of an interval	a
Class mark (class midpoint)	X
Random error or residual in linear regression	ϵ

REFERENCES

A BRIEF HISTORY OF STATISTICS (SELECTED TOPICS). ALPHA Seminar. 2017, 29 Ago [citado 08/07/2024]. Disponible en: http://homepage.divms.uiowa.edu/~dzimmer/alphaseminar/Statistics-history.pdf

Barrios JA. KARL PEARSON, CREADOR DE LA BIOMETRIA, PRECURSOR DE LA ESTADISTICA CONTEMPORANEA. Cizur Menor, España: Editorial Aranzadi; 2023.

Betanzos FG, de León M del CEP, López JKC. ESTADÍSTICA APLICADA EN PSICOLOGÍA Y CIENCIAS DE LA SALUD. Hipodromo: El Manual Moderno, Editorial; 2017.

Daniel WW. BIOESTADISTICA: BASE PARA EL ANALISIS DE LAS CIENCIAS DE LA SALUD. Editorial Limusa S.A. De C.V.; 2002.

De la Fuente Solana Joan Guàrdia Olmos EI. HISTORIA DE LA ESTADÍSTICA: UN EJERCICIO PARA ENTENDER EL MUNDO ACTUAL. España: Universidad de Barcelona; 2021.

De Mora Charles MAGVMS. HISTORIA DE LA PROBABILIDAD Y LA ESTADÍSTICA. España: UNED - Universidad Nacional de Educación a Distancia; 2018.

Del Cerro JS. HISTORIA DE LA PROBABILIDAD Y DE LA ESTADÍSTICA XII. Cizur Menor, España: Editorial Aranzadi; 2023.

Fernández JS-L. HISTORIA DE LA ESTADÍSTICA COMO CIENCIA EN ESPAÑA, (1500-1900). España: Instituto Nacional de Estadística; 1975.

Koren J. THE HISTORY OF STATISTICS, THEIR DEVELOPMENT AND PROGRESS IN MANY COUNTRIES: IN MEMOIRS TO COMMEMORATE THE SEVENTY-FIFTH ANNIVERSARY OF THE AMERICAN STATISTICAL ASSOCIATION. USA: American Statistical Association; 1918.

Lopez FJB. BIOESTADÍSTICA. España: Ediciones Paraninfo, S.A; 2005.

Macho JMA. HISTORIA DE LA ESTADISTICA. España: UNED - Universidad Nacional de Educación a Distancia; 2006

Martín MA, Horna O, Nedel F, Navarro A. FUNDAMENTOS DE ESTADISTICA EN CIENCIAS DE LA SALUD. Barcelona: Servei de Publicacions de la Universitat Autònoma de Barcelona; 2010.

Miranda RHC. BIOESTADÍSTICA APLICADA A LAS CIENCIAS DE LA SALUD. España: Formación Alcalá, S.L.; 2017.

Stigler SM. THE HISTORY OF STATISTICS: THE MEASUREMENT OF UNCERTAINTY BEFORE 1900. Cambridge, Mass., Estados Unidos de América: Belknap Press; 1990.

Vilalta JS. MANUAL DE BIOESTADÍSTICA. Barcelona, Spain: Elsevier Masson; 2003.