Unveiling Statistics in Articles

No fuss, almost no formulas, no tons of jargon

Paulo Buchsbaum

ISBN: 978-65-00-73102-6
Cover: Mahii Creations (Fiverr)

Acknowledgments

Thanks to Helena, for the constant encouragement, to my daughter Dr. Diana Buchsbaum, for her opinions, evaluations and comments and for being the person she is, to Marcio Caio, my brother, always supporting me with gestures and actions, besides a thorough review.

To my friend, Dr. Salim Kanaan, for providing the initial idea, in addition to always providing valuable criticism and suggestions.

To Daniel Tausk, PhD in Mathematics and professor at University of São Paulo (USP), for his valuable and detailed observations on the book that contributed to its final writing.

To my friend Luan Lins, almost a dentist, who, when I helped him deal with the statistics linked to his final work for his graduation, motivated me to turn my notes into a book.

"Tell me and I will forget,
show me and I may remember,
involve me and I will understand"
variation of a quote from Confucius

Foreword

Statistics can be considered a branch of applied mathematics, involved in a wide range of phenomena of many areas of knowledge. It represents a valuable work tool but ignored by many, due to the widespread aversion to numbers in our society.

The idea of this book is not to make you an expert in statistics. It will teach you to be careful when reading a scientific article or a journalistic writing that uses statistical data, in order to have more evaluation criteria.

The book was planned to introduce the statistical concepts (without the use of formulas) of greater relevance for the understanding of scientific articles.

In addition, the book describes and exemplifies several pitfalls, inside and outside statistics, that can undermine an article, encouraging the reader to have a more critical view.

Statistics is a very useful tool when used correctly and can be potentially dangerous in the wrong hands.

There is an interesting sentence said by a Swedish mathematician and writer Andrejs Dunkels: *"It's easy to lie with statistics, but it's hard to tell the truth without them"*

If you want to uncomplicate statistics and understand the numbers behind scientific articles, with greater accuracy, this is the book you need to read.

I can only congratulate Paulo Buchsbaum for this book, which serves as a great facilitator for all those who consume or intend to consume scientific content, whatever the area (medicine, psychology, nutrition, sociology, politics and so forth).

Prof. Salim Kanaan

Physician, professor at Fluminense Federal University (UFF), Master of Science in biophysics from Federal University of Rio de Janeiro (UFRJ), researcher and author of books such as "*Clinical Biochemistry*" and "*Acute Myocardial Infarction*"

Contents

1) Introduction

Most scientific articles (excluding descriptive, philosophical, or case studies), including articles in the areas of humanities and biomedical, need to use statistics to support the presented hypotheses.

From there, many readers of scientific research, especially those far from academia, are not familiar with the necessary jargon and concepts to get an overview of a scientific article and the degree of evidence that the article brings (or not).

This is what motivated me to write this book: to provide the reader with necessary and sufficient statistical information for reading articles, seeking an intuitive view, without drowning him in concepts, terminology, or formulas.

The fact is that I have not found any books, videos, or materials that come close to the content of this book.

Although there is much basic statistics content, it falls into two categories: one that discusses it in a verbose and casual manner, without a specific focus and blending entertainment with science; and one that shows their fundamental concepts for general use, but still involves the use of calculations and formulas.

The focus of the book is for the reader comes to understand the statistical part of an article, thereby fully understanding the conclusions, interpreting the results, and engaging with the discussion presented in the article. It is clearly not the intention here to give the reader the ability to reproduce a study in all its details.

The prerequisites for understanding this book are really minimal: will, persistence, interest, and nothing more than the four operations, decimals, and percentages.

Other than the concept of averages, formulas are not necessary because understanding the basics of article statistics does not require them.

Nowadays, formulas and heuristics are no longer so necessary, often not even for researchers, who do need to have someone on the team with training or statistical knowledge to select the models and parameters and to interpret the data and results in more depth. However, specialized packages and programs are usually used when applying formulas and models.

The central idea was to prepare the book so that its content was as simple, direct, and colloquial as possible, making every effort possible and almost impossible so that the reader can understand the most relevant concepts, which, when explained in the right and uncomplicated way, would become easy.

The motto here is to write to communicate without wanting to prove knowledge or dive into unnecessary details. So, here and there, I have preferred to sacrifice accuracy for clarity to achieve the goal of encouraging readers to meet this challenge.

In chapter 2, the book contains a very brief introduction about research and how to find scientific articles.

Chapter 3 introduces some basic concepts, plus graphs, means and dispersions.

Chapter 4 covers probability, correlation, regression and distributions.

Chapter 5 details the statistical concepts used in research.

Chapter 6 describes the most common types of scientific research.

Chapter 7 reveals some biases that can undermine an article with examples.

Chapter 8 quotes some small excerpts from scientific articles, exemplifying and commenting on how each one presents statistical data.

The book ends with a summary in six questions to show the whole in a few words, followed by an epilogue that motivates readers to give due importance to scientific research, but always with a questioning spirit.

Great read for anyone willing to navigate through these pages!

2) Overview of Scientific Research

Scientific research, in theory, follows a systematic process using a certain methodology, which is usually described, and generally aims in some way to contribute to human knowledge.

Research is often published in specialized scientific journals, which require the work to be reviewed by other scientists of a similar specialty. This process is called *peer review* because the reviewer is almost always a researcher like the authors are.

In most cases, this work is unpaid and voluntary. The personal benefit for reviewers is that being on the review team of a qualified journal brings prestige to those involved.

The review can be open (authors and reviewers know the identity of the other party), single-blind (authors do not have access to reviewers, but reviewers have access to the authors' names), or double-blind (the reviewer has no knowledge of who the authors of the work are, and vice versa).

As for the reader, he can evaluate the prestige of researchers (https://www.scopus.com/freelookup/form/author.uri) and journals, (https://www.scimagojr.com/journalsearch.php) which allows him to have an idea of the relevance of the work, although this does not necessarily guarantee that it is good.

Several portals index scientific articles. The most famous is Google Scholar (https://scholar.google.com), which allows the user to search for articles using keywords and to include in the query words in quotation marks to force the article to contain the words in a particular order, as in *Google Search*.

Example: *"glucose level"*

There is also the intitle *prefix:* used to search for articles in which the quoted words occur in the title and can be used with or without quotation marks.

Example: *intitle:glucose intitle:diabetes* or *intitle:"hearing happens"*

The results are ordered by relevance, a criterion that is not fully transparent. Google Scholar includes for each article a count of citations to the article by other sources, which tends to give some measure of its scholarly relevance.

Much scientific research depends on validating the claims made through statistical tools, which may involve existing data or new data to be produced by the work itself.

With the introduction of the Internet, research began to be published on certain portals without the need for peer review. The lack of peer review does not necessarily imply that it is low-quality

material, although the peer review process increases the chance that the research will be higher quality.

Nowadays, everyone has access to most published articles; however, access in many cases is limited to the article abstract, and paying a fee or having affiliation with an academic institution that subscribes to the content is necessary to have access to the complete article.

3) Descriptive Statistics

Descriptive statistics refers to techniques for describing data, synthesizing data from metrics that help us understand them, or expressing them in the form of tables or various graphs.

3.1) Variables, Populations, and Samples

When conducting an objective study, we almost always have something we want to evaluate (cures, effects, deaths, religion, votes, or similar), which can be called an *event, variable, factor, attribute,* or *feature*. These items can be generically defined as anything that can be counted, measured, or classified relative to a condition, fact, object, person, animal, etc.

Factors can involve quantitative or numerical measures (such as weight in kg) or value ranges (income ranges to estimate income tax rate). Quantitative measurements can be discrete (integers, such as number of surgeries performed, since no one does half a surgery) or continuous (real numbers, such as blood glucose level or percentage of oxygen in the air), which can be any value within a reasonable range.

Factors can also be categorical (described by categories) or qualitative. In these cases, the terms *attribute* and *characteristic* are used more often. Qualitative factors can be totally subjective (such as level of happiness), moderately subjective (such as race), or more objective (such as profession). It may also involve an ordered attribute, such as obesity qualification (morbid obesity, overweight, and so on), or others that are not ordered, such as profession or sex.

An event may involve the cure or onset of disease *X*, people taking medication *Y*, blood pressure, blood pressure range (such as low, normal, high), age, age group, death, age, weight, and so forth. In general, an event involves some context of time (time or measurement interval) and place (such as city, state, country).

Any study is carried out within a context called *population*, which refers to the total number of items (animals, people, objects) that are subject to the studied event, to which filters or conditions may apply (which may involve characteristics such as race, age, weight, and comorbidities) in certain places and times.

For example, a researcher decides to study the effect of medicine *X* on only obese and diabetic men. Thus, out of all people, the study population includes people with an obesity filter, with the presence of diabetes, and who are men.

Generally, researching the entire population is impossible, even filtering by certain attributes, so a *sample* is selected, which

represents a subset of items (in this case, people) within the desired universe.

Before continuing to talk about samples, it is necessary to distinguish between two concepts: *incidence* and *prevalence*.

In the case of a disease, condition, or characteristic, the ratio between new events within a time slice (1 year, for example) and the corresponding population is called the *incidence,* whereas the *prevalence* is also a ratio but refers to the total percentage of occurrence of the given factor at a moment in time. Even the prevalence tends to change over time because the incidence can increase or decrease with the passage of time.

For example, twins make up about 1.2% of the population (prevalence), which is around 95 million twins for the population of 7.9 billion. Because the incidence has been higher, it is estimated at around 1 for every 40 live births (i.e., 2.5%, about twice the prevalence).

The most important types of samples are *random sampling* (in which individuals are randomly selected by some type of chance, such as address or document number), *clustered sampling* (in which locations within the universe are selected, for reasons of time or cost, and then random sampling is done), or *stratified sampling*, which is explained below.

In stratified sampling, the target population is divided into *strata*, which represent criteria such as age, weight, and comorbidities. From there, random people are chosen proportionally to the prevalence of each stratum in the population. As an example, let us say one wants to research four age groups and three degrees of obesity based on an estimate of the prevalence percentage of each age group in the population; thus, we have a total of 4 x 3 = 12 strata of people, each representing a combination of age group and an obesity classification.

Each piece of data can be presented in absolute form (quantity) or as a percentage relative to the corresponding total.

A variable can then be expressed in a set of data that represents a given sample; in this case, the number of data referring to a variable corresponds to the sample size. Thus, if we collect blood pressure data from 200 people, we will have 200 data points in total. A variable can even refer to the entire target population, which is generally not possible except in very limited circumstances.

3.2) Graphs in Statistics

Graphs and *charts* represent visual ways of displaying data, which greatly facilitates the visualization and interpretation of the relationship between these data when compared to text.

In the case of graphics in statistics, we are usually interested in visualizing the relationship between two or more variables that help to describe a generic item, which can be a person, object, country, and so on.

In this book, we will limit ourselves to just two variables, which are two-dimensional graphics, precisely referring to two variables or dimensions. This is the most used chart type.

For example, if we study people and we are working with the variables age and number of cars that each one owns, we can assume, for our example, three people: José, 23 years old and no car; Maria, 34 years old and one car; and Pedro, 56 years old and two cars. So, we have three pairs of values, which can be grouped between parentheses [(23, 0), (34, 1), and (56, 2)] and are visualized in Figure 3.1 below.

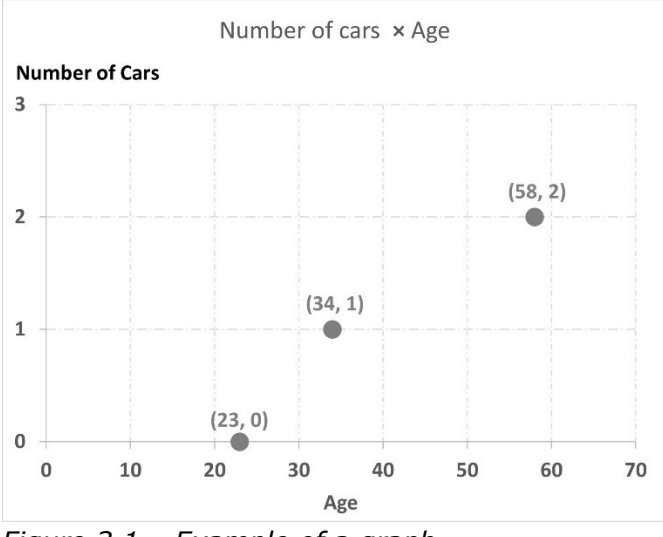


Figure 3.1 – Example of a graph

The most common way to proceed is to display each variable on an axis, each represented by a straight line. There are two perpendicular axes (they form a 90º angle). One axis is horizontal and represents the first desired variable, and the second axis is vertical and represents the other variable. In general, the independent variable (the "cause" of some event) is represented horizontally and the dependent variable (the event that is supposedly the "consequence") vertically, as shown in Figure 3.1 above. If there is no supposed causal relationship, the

choice of which variable goes on each axis is arbitrary and depends on taste and tradition.

The horizontal axis is generically referred to as *X* but is usually named after the associated variable (e.g., age, temperature, salinity). The vertical axis is generically referred to as *Y* but is generally named after the associated variable (e.g., number of cars, blood pressure, flow).

Note that conventionally a graph can be referred to using "*Y* axis × *X* axis." For example, consider "Number of cars × Age," where × means *versus* (number of cars versus age), as if it were a comparison.

When a specific point is referred to, it is represented by a pair of values. For example, the second point in Figure 3.1 above refers to the pair (34, 1), which in this case represents the point that results from crossing the vertical that comes out above the "Age" axis at the value 34, between 30 and 40, with the horizontal to the "No. of Cars" axis, which comes out to the right of the value 1.

The axes must be given a scale, which may be represented by ticks, which are small lines cutting each axis perpendicularly (which in the figure above appear below the horizontal axis and to the left of the vertical axis). The optional horizontal and vertical line set within the chart seen in Figure 3.1 above is called a grid and helps to visualize the values.

In *linear charts* (by far the most common type of chart), the ticks are equally spaced on the axes in terms of values. Thus, age is represented by values every 10 years, receiving ticks at 0, 10, 20, 30, and so on.

In the example given, it seems that the older the person, the more cars he or she tends to own, but this is obviously a misleading conclusion because only three people were considered.

For use in statistics, four types of graphs are the most common, as follows:

The graph explained above is the most used and can be represented by points (as in Figure 3.1 above) or by a line. In Excel®, this is called a *scatter plot.*

The *legend* is a box that explains some elements of the graphic and is not always necessary. In Figure 3.2 below, the legend is omitted because the information is in the chart itself.

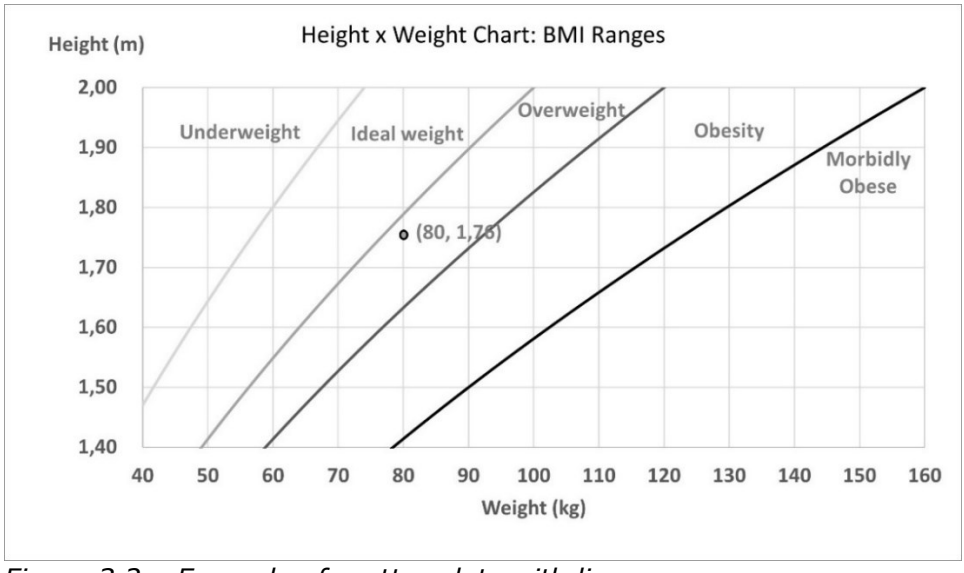


Figure 3.2 – Example of scatter plots with lines

Note the curves separating the various weight categories according to BMI ranges (*Body Mass Index = weight / height*^{2}). The visible point in the figure (represented by a small black circle) is the intersection of 80 kg and 1.76 m.

In this case, the choice of the variable associated with the axes can be anyone, because weight does not determine height, nor does height determine weight (although it helps). This chart has no legend because the explanation is placed inside the chart.

If the representation of the relationship between two variables is given by isolated points, one can draw a curve that *fits* the data points in the best way, in a mathematical process called *regression*, which will be discussed a little later.

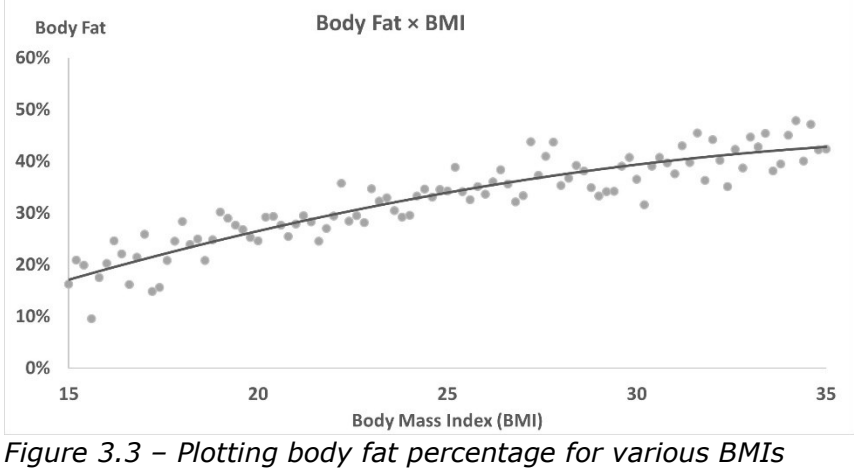


Figure 3.3 – Plotting body fat percentage for various BMIs

In Figure 3.3 above, BMI data and body fat percentages (measured with a specialized device) of several people were plotted, represented

by scattered points in the graph. Using regression techniques, a curve was drawn through the middle of the points, which expresses body fat as a function of BMI.

The second type of chart used in statistics is the *bar chart,* which is often used for qualitative data. In this case, the horizontal axis represents a nonnumerical category, such as the day of the week, as seen in Figure 3.4 below, and the vertical axis can represent any kind of value. In the case of Figure 3.4, it is the *probability* of birth per day of the week. (*Probability* will be discussed in the next section.)

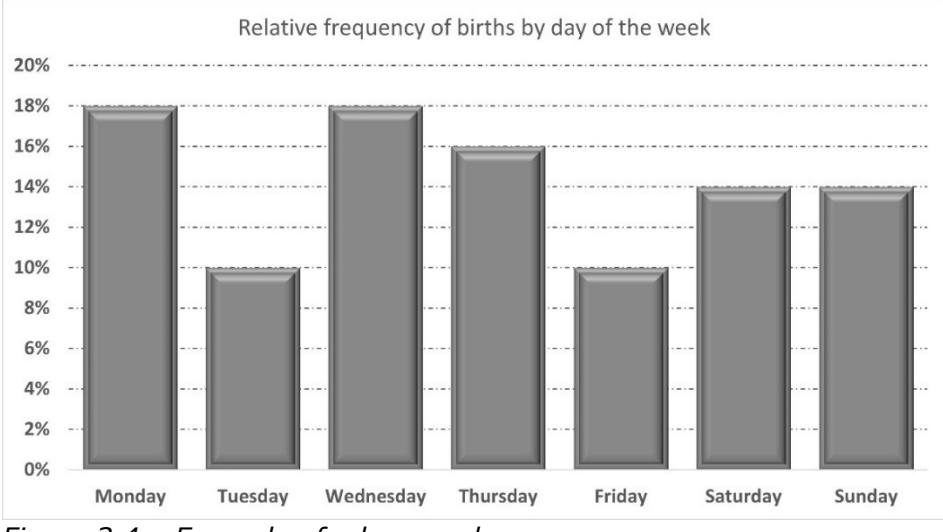
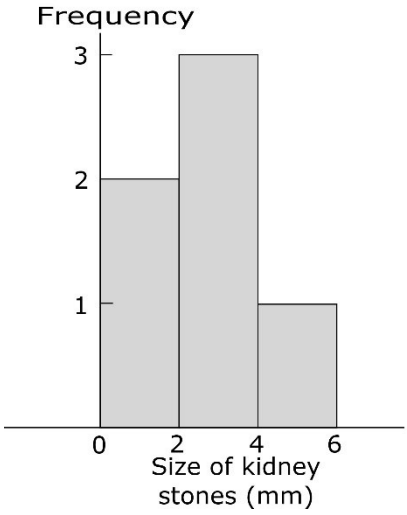


Figure 3.4 – Example of a bar graph

The third type of graph is the *histogram*, used to represent *ranges* (each representing a *bin)* of values of a variable (such as grade ranges or blood cholesterol levels) on the horizontal axis, whereas the frequency (count) of values within each range appears on the vertical axis (or the probability of a value being in this range). The bars are glued together, with intervals of the same width.

Usually, the range includes the lower end but not the upper end, as can be seen in Figure 3.5 below.



Stones (mm)
1,0
1,5
2,0
3,2
3,9
5,1

Figure 3.5 – Example of a histogram

In Figure 3.5 above, the bars are associated with bands or intervals of 2 mm width. Thus, the 1 and 1.5 mm stones are on the first bar (from 0 to 2), whereas the 2, 3.2, and 3.9 mm stones are on the second bar (from 2 to 4), and finally the 5.1 stone is on the third bar. Note that the 2 mm stone was in the second bar and not the first because the lower end of the range on the right generally prevails.

The fourth and final chart type is almost self-explanatory. This is the *pie chart,* which helps to convey the impact of each value option on categorical (qualitative) data.

In the example in Figure 3.6 below, the legend on the right communicates what each area represents without polluting the inside of the pie with too much text.

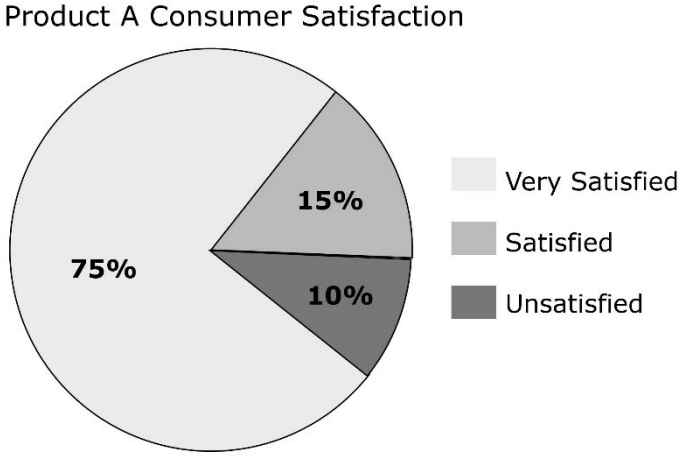


Figure 3.6 – Example of a pie chart

3.3) Measures of Central Tendency

In the case of quantitative data, certain measures express values containing information about the dataset. These measures are called *metrics* or *parameters*.

A *measure of central tendency* estimates a single value that represents the entire set, in terms of the value's size.

Of the several measures of central tendency, the most important is the *arithmetic mean*—or simply the *mean* (more often used in statistics) or *average (*general usage and more colloquial), which simply consists of adding all the values and dividing by the number of elements.

Example: If the test scores in a class are 8, 7, and 3, the average is simply (8 + 7 + 3) / 3, which is 6.

Note that the mean is the most representative value of the set of values, displaying the typical "size" of the data.

The mean has an interesting property: if we add all the deviations from the mean for all the data in the set (recalling that values greater than the mean have a positive deviation and values less than the mean have a negative deviation), the sum will always be equal to 0, which demonstrates that the mean is in fact exactly in the middle of the data.

This property helps improve intuition about the meaning of the mean and is easy to visualize.

Assume an average of 4 values. It could be 8, 10, or as many values as one wants; this is just an example to facilitate understanding.

mean = (1st value + 2nd value + 3rd value + 4th value) / 4

Hence, we find that the deviations of each value from the mean is equal to 0:

(1st value − mean) + (2nd value − mean) + (3rd value − mean) + (4th value − mean) = 0

Why? Grouping the data, the mean subtracts four times:

(1st value + 2nd value + 3rd value + 4th value) − mean × 4 = 0

Moving *− mean × 4* to the right as positive becomes:

(1st value + 2nd value + 3rd value + 4th value) = mean × 4

Which finally comes back to the definition of mean as given above, reversing the sides and then passing *× 4* to the other side by dividing:

Mean = (1st value + 2nd value + 3rd value + 4th value) / 4

Another way to express data is through a dataset, where each data item can occur more than once or have an associated weight. This weight, especially if it means data repetition, can be called the *frequency*. In this case, the mean is called a *weighted average*.

This mean is calculated in the same way if each data point is counted in the sum as many times as its frequency, and at the end,

this sum is divided by the total frequency, which represents the real number of data points, counting the repetitions.

A simpler and equivalent way is merely to multiply each piece of data by its frequency (which corresponds to adding each data as many times as its frequency) and obtain this sum for all the data, dividing at the end by the sum of the frequencies or weights.

Example: Suppose the final passing grade is 5 and a student gets the following grades:

First grade: 1 weighted 1
Second grade: 1 weighted 1
Third grade: 10 weighted 2

Did this student move on to the next level? If there were no weights (or if all weights were 1), clearly not, because we would have 1 + 1 + 10 = 12, which averages 12 / 3 = 4.

However, here we do 1 + 1 + 10 + 10, as if the student had scored 2 times 10, because the third evaluation has a weight of 2, which in practice is equivalent to multiplying 10 by the weight (2) that corresponds to 1 + 1 + 10 × 2 = 22. Because the sum of the weights is 4, this results in 22 / 4 = 5.5, and this person would move on to the next level!

Another type of metric like the mean is the *median*, which consists of taking the data point from the middle of an ordered series of data. Thus, half of the values are equal to or less than this data point, and half of the values are equal to or greater.

The greatest advantage of the median is that, by nature, it disregards *outliers,* that is, values outside of the expected range. For example, when measuring the mean height of 11 people, if all fall between 160 cm and 180 cm except for a 120 cm little person, this greatly distorts the mean, which does not happen if we take the median.

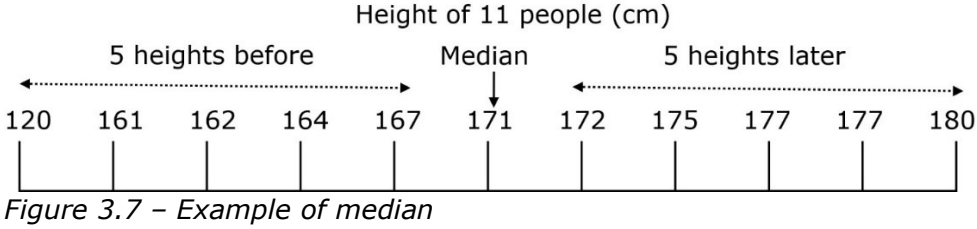


Figure 3.7 – Example of median

In this example, the height of the little person is an atypical point. After all, little people are not 10% of the population. The average, which uses this height, is correct, but it does not represent a typical value.

If so, why is not the use of the median as common as the use of the mean? It is less common because the mean is more appropriate from a mathematical point of view, due to its properties.

Finally, we have the *mode*, which represents the interval or value that occurs most often in a certain sample or population. It is a useful metric to visualize because it shows the peaks of the curve.

For example, in Figure 3.8 below, a rain chart for 1 year is shown for a hypothetical location where it rains every day of the year. The months appear on the horizontal axis, where the label marks the beginning of each month, and the amount of rain that fell each day appears on the vertical axis, in mm (height of water that rain leaves in an empty box). This exotic curve has two modes: on May 1st and October 1st, it rained 7 mm, the record for the year.

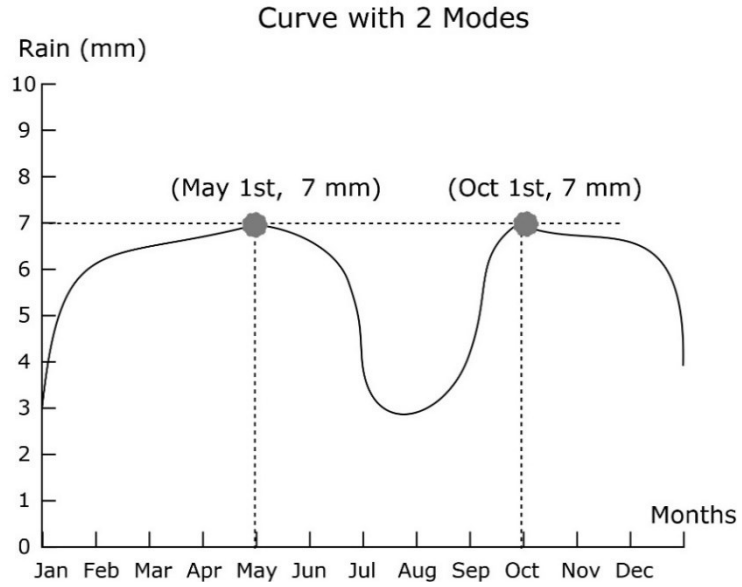


Figure 3.8 – Example of mode

The book between pages 19 and 108 is not shown here

Bibliography and References

Books

BUSSAB, WO; MORETTIN, PA *Basic Statistics* , 9th ed. São Paulo: Saraiva, 2017

DIEZ, D.; ÇETINKAYA-RUNDEL M.; BARR, OpenIntro Statistics CD , 4th ed., Boston: OpenIntro, 2022.

MANN, PS *Introductory Statistics* , 10th ed., Hoboken: Wiley, 2020.

MCCLAVE, J.; SINCICH, T., *Statistics* , 13th ed, Boston: Pearson, 2018

RUMSEY, DJ *Statistics For Dummies* , 2nd ed. Rio de Janeiro: Alta Books, 2019

Articles

MEHRAN, F. Sample size and margin of error, *International Labor Organization (ILO)*, Geneva, 2014.

SASHEGYI, A.; FERRY, D. On the Interpretation of the Hazard Ratio and Communication of Survival Benefit. The Oncologist, Oxônia, v. 22 n. 4, p. 484-486, 2017.

SELLA, F.; RAZ, G.; KADOSH, R. C. When randomisation is not good enough: Matching groups in intervention studies. Psychonomic Bulletin & Review, Berlin, v. 28, p. 2085-2093, 2021.

WHAKATUTUKI, H. Technical paper - empirical equations for IVS relative margin of error. *New Zealand Government*, 2015.

WHAKATUTUKI, H. International Visitor Survey Relative Margin of Error Empirical Equation. *New Zealand Government*, 2017.

YOUNG, H. L. Strengths and Limitations of Meta-Analysis. *The Korean Journal of Medicine*, Seoul, v. 4, n. 5 p. 391-395, 2019.

Other References

BASIC & Clinical Biostatistics Glossary 4th ed. *DoctorLib.* Available at https://doctorlib.info/medical/biostatistics/14.html. Accessed May 16, 2023.

DHAND, N. Demystifying statistics: Estimating sample size for a survey. *Statulator*, Sydney, 2015. Available at https://www.statulator.com/blog /demystifying-statistics-estimating-sample-size-for-a-survey. Accessed May 16, 2023.

FROST, J. Confidence Intervals: Interpreting, Finding & Formulas. *Statistics by Jim*, State College, PA. Available at

https://statisticsbyjim.com/hypothesis-testing/confidence-interval. Accessed May 23, 2023.

MESTER, T. Statistical Bias Types explained – part 1 (with examples) , *Data36*, Budapest, 2022. Available at https://data36.com/statistical-bias-types-explained. Accessed May 16, 2023.

MESTER, T. Statistical Bias Types explained – part 2 (with examples), *Data36,* Budapest, 20172. Available at https://data36.com/statistical-bias-types-examples-part2/. Accessed May 16, 2023.

MORGENSTERN, J. Bias in medical research, *First10EM*, Toronto, 2021. Available at https://first10em.com/bias. Accessed May 16, 2023.

MULLER, DA Is Most Published Research Wrong? Verisatum – YouTube, Los Angeles, 2016. Available at https://www.youtube.com /watch?v=42QuXLucH3Q. Accessed May 16, 2023.

SIMPSON'S paradox, Wikipedia, San Francisco. Available at https://www.wikiwand.com/en/Simpson_Paradox. Accessed May 16, 2023.

TYPES of experimental studies, Eupati Open Classroom, Bern. Available at https://learning.eupati.eu/mod/book/view.php?id=665. Accessed May 16, 2023.

Links cited in the text

Correlation and causality. *Khan Academy*, Mountain View. Available at https://www.youtube.com/watch?v=ROpbdO-gRUo. Accessed May 16, 2023.

Google Scholar. *Google*, Mountain View. Available at https://scholar.google.com.br. Accessed May 16, 2023.

How to Determine Sample Size from G*Power. *Statistics Solutions*, Palm Harbor FL. Available at https://www.statisticssolutions.com/how-to-determine-sample-size-from-gpower. Acessed May, 20, 2023.

Nurses' Health Studies. *Harvard – School of Public Health*, Boston. Available at https://www.hsph.harvard.edu/nutritionsource/nurses-health-study. Acessed May, 20, 2023

Scimago Journal & Country Rank. *Scimaco*, Grenada. Available at https://www.scimagojr.com/journalsearch.php. Accessed May 16, 2023.

Search for an author profile. *Scopus*, Amsterdam, Netherlands.

Available at https://www.scopus.com/freelookup/form/author.uri. Accessed May 16, 2023.

Stanford Profiles - John PA Ioannidis. *Stanford University*, Stanford. Available at https://profiles.stanford.edu/john-ioannidis. Accessed May 16, 2023.

Articles cited in the text

ANTONINI, M. et. al. Does Pink October really impact breast cancer screening? *Public Health in Practice*, Oxford, v. 4, set. 2022.

BOKHARI, SAH Non-surgical periodontal therapy reduces coronary heart disease risk markers: a randomized controlled trial. *Journal of Clinical Periodontology* , Hoboken, v. 39 no. 11 p.m. 1065-1074, 2012.

BOUVARD, V. et. al . Carcinogenicity of consumption of red and processed meat. *The Lancet Oncology*, v. 16, n.16, p. 1599-1600, dec. 2021.

BRIASOULIS, A.; MD, Vikram AGARWAL, V.; Franz H. MESSERLI, FH Alcohol Consumption and the Risk of Hypertension in Men and Women: A Systematic Review and Meta-Analysis. *American Journal of Hypertension* , Oxford, v. 14, no. 11, p. 792-798, 2012

CHOW, CK Effect of Lifestyle-Focused Text Messaging on Risk Factor Modification in Patients With Coronary Heart Disease. A Randomized Clinical Trial. *Jama* , Chicago v. 314, no. 12 p. 1255-1263

DOLL, R.; HILL, AB Smoking and Carcinoma of the Lung. *British Medical Journal* , London, v. 2 n. 4682, 1950.

IANGARI, SH Coronary Heart Disease and ABO Blood Group in Diabetic Women: A Case-Control Study. *Nature* , London, Scientific Reports, 2019.

IOANNIDIS, JPA Why Most Published Research Findings Are False. *PLOS Medicine* , San Francisco, v. 2 n. 8, 2005

IOANNIDIS, JPA et. al. A Manifesto for Reproducible Science. *Nature Human Behaviour,* London, v. 1 n. 21, 2017

KAPTCHUK, TJ; MILLER, GM Placebo Effects in Medicine. *New England Journal of Medicine*,
Boston, v.373, n. 1, p. 8-9, 2015.

NGAHANE, B.H.M.; EKOBO, H. A; KUABAN. Prevalence and determinants of cigarette smoking among college students: a cross-sectional study in Douala, Cameroon. *Archives of Public Health*, Berlim, v. 73, n. 47, dez. 2015.

PONIKOWSKI, P. Ferric carboxymaltose for iron deficiency at discharge after acute heart failure: a multicentre, double-blind, randomized, controlled trial, *The Lancet* , London, v. 396, no. 10266, p. 1895-1904, 2020.

SERDAR, C. C. et. al. Sample size, power and effect size revisited: simplified and practical approaches in pre-clinical, clinical and laboratory studies. *Biochemia Medica*, Zagreb, v. 31 no. 1, 2021

WHITEMAN, DC et al. Cancers in Australia in 2010 attributable to modifiable factors: summary and conclusions. *Australian and New Zealand Journal of Public Health* , v. 39, no. 5, p. 477-484, 2015.