Chapter

Descriptive Statistics

Hazhar Talaat Abubaker Blbas

Abstract

Descriptive statistics is a branch of statistics that deals with summarizing and describing the main features of a dataset. This chapter will cover the value of statistics, how data analysis occurs in scientific study, the distinction between a sample and the population, the different types of variables, sampling techniques, measures of central tendency, and measures of dispersion. The summary of descriptive statistics gives a succinct overview of various metrics and visual representations, enabling researchers and analysts to learn more about the features of the dataset and draw accurate conclusions.

Keywords: mean, median, mode, standard deviation, coefficient of variation, probability sampling, non-probability sampling

1. Introduction

Descriptive statistics involve summarizing and describing data using numerical measures and graphical representations. It provides a concise and meaningful way to understand and communicate the main characteristics of a dataset. This introduction explores the basics of descriptive statistics, including measures of central tendency, measures of dispersion, and graphical representations. By examining these statistical tools, we can gain insights into the patterns, variability, and distribution of data, allowing us to make informed interpretations and draw meaningful conclusions.

2. What is the process of analyzing data in statistics?

Statistics is the science of collecting, organizing, analyzing, and interpreting data in order to make decisions as shown in **Figure 1**.

3. Sample and population

A population is the collection of all outcomes, responses, measurements, or counts that are of interest since sample is a subset of a population as shown in **Figure 2** [1–3].

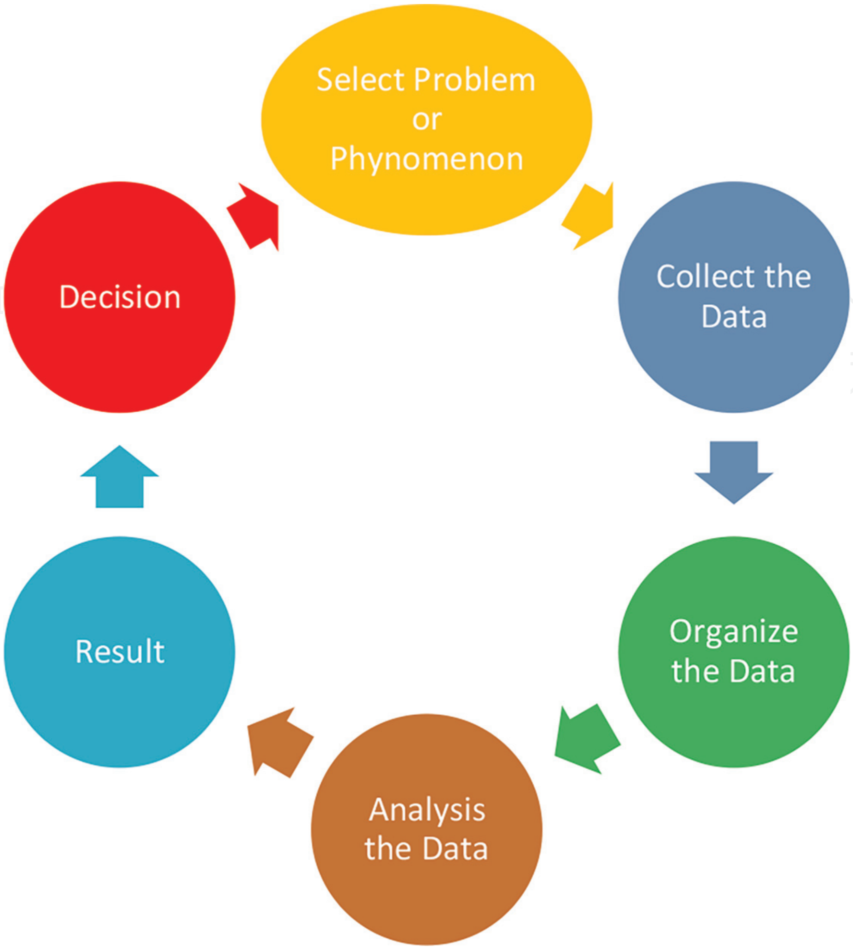

Figure 1.
Process of data analysis in scientific research. Source: Author has been created as a new work.

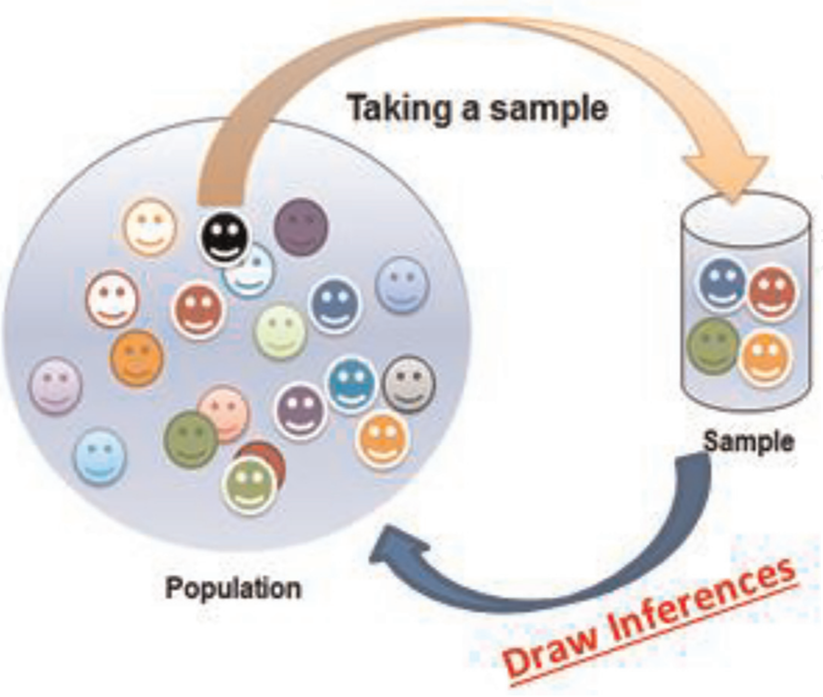

Figure 2.
Difference between sample and population. Source: Author has been created as a new work.

4. Type of variables

Variable is a characteristic that can assume different values and alphabetic. There are two common types of variables, such as quantitative variables and qualitative variables, as shown in **Figure 3** [4, 5].

1. Quantitative variables (numerical variables) are variables that represent measurable quantities or amounts. They can be further classified into two types:

    i. Discrete variables: Discrete variables are numerical variables that can only take on specific, separate values. These values are typically whole numbers or counts and cannot be subdivided further. Examples of discrete variables include the number of children in a family, the number of customers in a store, or the number of items sold.

    ii. Continuous variables: Continuous variables are numerical variables that can take on any value within a certain range. They can be measured with a high degree of precision and can have infinite possible values between any two points. Examples of continuous variables include height, weight, temperature, and income.

2. Qualitative variables (categorical variable): This type of variable represents data that can be divided into distinct categories or groups. Examples include gender, ethnicity, marital status, and level of education.

    i. Nominal variables: Nominal variables are categorical variables that represent data with no ranking. Examples of nominal variables include gender (male/female), ethnicity (Asian, African, European, etc.), marital status (single, married, divorced), and eye color (blue, brown, green).

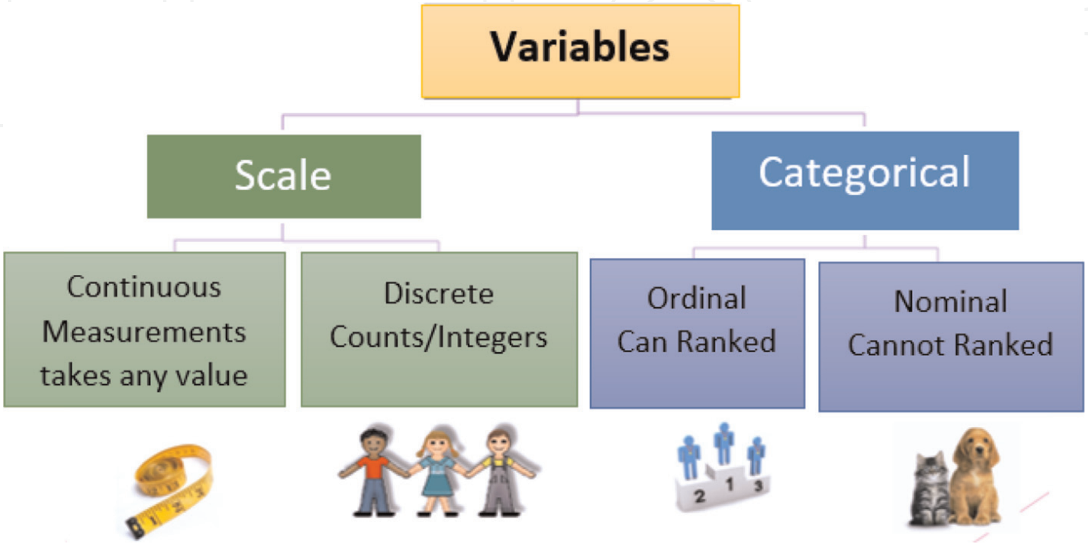

Figure 3.
Type of variables. Source: Author has been created as a new work.

ii. Ordinal variables: Ordinal variables represent data that has a natural order or ranking, but the differences between the categories may not be consistent or measurable. The categories can be ranked or ordered based on some criterion, but the magnitude of the difference between categories is not known. Examples of ordinal variables include satisfaction levels (very satisfied, satisfied, neutral, dissatisfied, very dissatisfied), educational attainment (high school diploma, bachelor's degree, master's degree), and survey responses using Likert scales ("strongly agree," "agree," "neutral," "disagree," "strongly disagree").

5. Sampling plan

Once the target population has been identified, next the sampling plan must be devised. Goal: Randomly select a small percent of the population that will in turn represent the ideas of the population as a whole. There are two general types of sampling techniques [1, 2, 5]:

5.1 Probability (random) sampling

All members of the population must be specified prior to drawing the sample and each member of the population has equal probability of being chosen or included in the sample. There are four common types of Probability (Random) Sampling:

5.1.1 Simple random sampling

Simple random sampling is a statistical sampling technique in which each member of a population has an equal probability of being selected to be part of the sample. The selection process is conducted randomly, without any bias or preference toward certain individuals or elements in the population.

A researcher wants to conduct a survey to understand the opinions of students at a university regarding a new policy. The university has a total population of 1500 students. For example, a researcher wants to select 100 out of 1500 students as a sample. Put a unique identifier to each of the students such as a student ID number. Then, randomly select the 100 students as a sample like a lottery game.

5.1.2 Systematic sampling

Systematic sampling is a statistical sampling technique that involves selecting every kth element from a population, where k is a predetermined interval. It is similar to simple random sampling but incorporates a systematic approach to the selection process.

Depending on the previous example of simple random sampling, the researcher wants to select 100 students using systematic sampling. We will calculate the sampling interval, which divides the population size by the desired sample size to determine the sampling interval. In this case, the sampling interval would be 1000/100 = 10. Next, select a random starting point within the first k elements (in this case, the first 10 students). Next, starting from the random starting point, select every 10th student thereafter. So, you would select the 10th, 20th, 30th, and so on, until you reach the desired sample size.

5.1.3 Stratified sampling

Stratified sampling is a statistical sampling technique that involves dividing the population into two or more than two homogeneous groups. Then, randomly select the desire case in each groups using simple random sampling.

Depending on the previous example in simple random sampling, the researcher wants to select 100 students using stratified sampling.

First, students can be stratified based on their academic disciplines into four strata: statistics, accounting, business, and economics department.

1. Determine the sample size: Decide on the desired sample size for each stratum. Let us say you want to sample 25 students from each stratum (department), resulting in a total sample size of 100 students.

2. Divide the population into four strata: Categorize the students into the respective strata based on their academic disciplines. Each student should belong to only one stratum.

3. Determine the allocation: Calculate the proportionate allocation for each stratum by dividing the desired sample size for that stratum by the total sample size. In this case, since each stratum has the same desired sample size (25 students), the allocation would be 1/4 (25%) for each stratum.

4. Sample within each stratum: Perform simple random sampling within each stratum separately. Randomly select 25% (25 students) from the statistics stratum, 25% from the accounting stratum, 25% from business stratum, and 25% from the economics stratum.

5. Collect data: Once the samples are selected, collect the relevant data or information from the students in each stratum.

5.1.4 Cluster sampling

Cluster sampling: Cluster sampling involves dividing the population into clusters or groups, often based on geographical proximity, and randomly selecting entire more than one clusters as the sampling units. This technique is useful when it is impractical or costly to sample individuals individually, and it can provide cost and time efficiencies.

5.2 Nonprobability sampling

Every element in the population does not have an equal probability of being chosen. The process of inclusion in the sample is based on the judgment of the person selecting the sample. There are four common types of nonprobability sampling.

5.2.1 Judgment sampling

Purposive sampling: Purposive sampling, also known as judgmental or selective sampling, involves handpicking individuals based on specific criteria or the researcher’s judgment. This technique is often used in qualitative research or when a specific subgroup of the population is of particular interest. Purposive sampling allows the researcher to target individuals who possess the desired characteristics or have relevant experiences.

5.2.2 Convenience sampling

Convenience sampling: Convenience sampling involves selecting individuals who are easily accessible or readily available to the researcher. This method is convenient and often used in situations where time, cost, or accessibility is a constraint. However, convenience sampling can introduce bias, as the sample may not be representative of the entire population.

5.2.3 Quota sampling

Quota sampling: Quota sampling involves setting specific quotas or targets for certain characteristics or subgroups within the population. The researcher selects individuals to fulfill the predetermined quotas until they are satisfied with the sample composition. Quota sampling allows for control over sample proportions but does not involve random selection.

5.2.4 Snowball sampling

Snowball sampling: Snowball sampling is a technique where initial participants are selected, and then they help identify and recruit additional participants from their social networks. This method is useful when studying hard-to-reach or hidden populations. Snowball sampling relies on referrals and networks to expand the sample size.

6. Measures of central tendency

It is a statistical measure that represents information about the central or middle value of a dataset. The three common measures of central tendency are the mean, median, and mode [4–6].

1. Mean (average), is calculated by summing up all the values in a dataset and dividing by the number of values. It represents the balancing point of the dataset and is sensitive to outliers. Depending on 894 people from Kurdistan Region of Iraq, the average age of people for the survey about depression and anxiety during the outbreak of COVID-19 is 33 years [1].

$$\overline{X} = \frac{\sum X_i}{n} \tag{1}$$

   Example: Consider the following dataset of exam scores: 85, 90, 92, 88, 95. The mean is calculated as (85 + 90 + 92 + 88 + 95) / 5 = 90.

2. Median: The median is the middle value in a dataset when it is arranged in ascending or descending order. If there is an even number of values, the median

is the average of the two middle values. The median is less influenced by outliers compared to the mean.

Example: Using the same dataset of exam scores: 85, 90, 92, 88, 95. When arranged in ascending order, the middle value is 90. Therefore, the median is 90.

3. Mode: The mode represents the most frequently occurring value(s) in a dataset. It is the value that appears with the highest frequency. A dataset can have no mode (when all values occur equally) or multiple modes (when multiple values have the same highest frequency).

Example: Consider the following dataset of exam scores: 85, 90, 92, 88, 90. The mode is 90 because it appears twice, which is more frequently than any other value.

7. Measures of dispersion (variation)

Measures of dispersion (Variation), provide information about the spread or dispersion of data points around the central tendency. The first three main measures of dispersion including range, standard deviation, and variance, are used when we have the same unit of datasets but we can use coefficient of variation once we have different units of datasets [4–8].

1. Range (R): It is the difference between the maximum and minimum values in a dataset.

$$R = Highest\ value - Lowest\ value \tag{2}$$

Example: Consider the following dataset of exam scores: 85, 90, 92, 88, 95. The range is calculated as 95–85 = 10.

2. Variance (S^2): It measures the average squared deviation of each data point from the mean. It provides a more precise measure of dispersion by considering the differences between individual data points and the mean. However, it is in squared units and is sensitive to outliers.

$$S^2 = \frac{\sum \left(X_i - \overline{X}\right)^2}{n-1} \tag{3}$$

Example: Using the same dataset of exam scores: 85, 90, 92, 88, 95. The variance is calculated as follows:

- Calculate the mean: (85 + 90 + 92 + 88 + 95) / 5 = 90.

- Calculate the squared deviation for each data point from the mean: (85–90)^2, (90–90)^2, (92–90)^2, (88–90)^2, (95–90)^2.

- Calculate the average of these squared deviations: (25 + 0 + 4 + 4 + 25) / 5 = 12.8. Therefore, the variance is 12.8.

1. Standard Deviation (S): It is the square root of the variance. It is the most commonly used measure of dispersion as it is in the original units of the data, making it more interpretable. It provides a measure of how much the data deviates from the mean.

$$S = \sqrt{\frac{\sum \left(X_i - \overline{X}\right)^2}{n - 1}} \tag{4}$$

Example: Using the same dataset of exam scores: 85, 90, 92, 88, 95. The standard deviation is the square root of the variance calculated in the previous example, which is approximately 3.58.

2. The coefficient of variation (CV) is a relative measure of dispersion that expresses the standard deviation as a percentage of the mean. It is used to compare the variability of datasets with different means or scales. The formula for calculating the coefficient of variation is:

$$CV = \frac{S}{\overline{X}} * 100 \tag{5}$$

Here's an example to illustrate the calculation of the coefficient of variation. Consider two datasets representing the monthly sales of two stores:

Store A: Mean = $10,000, Standard Deviation = $2000.
Store B: Mean = $15,000, Standard Deviation = $3000

- CV foe the store A = (2000 / 10,000) * 100 = 20%

- CV foe the store B = (3000 / 15,000) * 100 = 20%

In this example, both stores have the same coefficient of variation of 20%. It indicates that the relative variability or dispersion of sales is the same for both stores, even though Store B has a higher mean and standard deviation compared to Store A.

A lower coefficient of variation indicates less variability relative to the mean, while a higher coefficient of variation suggests greater relative variability.

Additional information

ORCID account: https://orcid.org/my-orcid?orcid=0000-0002-5760-3019

Google Scholar Citation: https://scholar.google.com/citations?user=zl0ee JoAAAAJ&hl=en&authuser=1

Author details

Hazhar Talaat Abubaker Blbas
Department of Statistics, College of Administration and Economics, Salahaddin University, Erbil, Kurdistan Region, Iraq

*Address all correspondence to: hazhar.abubaker@su.edu.krd

References

[1] Aroian K, Uddin N, Blbas H. Longitudinal study of stress, social support, and depression in married Arab immigrant women. Health care for women international. Feb 1 2017;**38**(2): 100-117

[2] Rosner B. Fundamentals of biostatistics. Cengage Learning. 2015

[3] Bluman A. Elementary Statistics: A Step by Step Approach 9e. McGraw Hill; 2014

[4] Blbas H. Statistical analysis for the most influential reasons for divorce between men and women in Erbil-Iraq. International Journal. Malmö, Sweden. 2019

[5] Triola MF, Iossi L. Essentials of Statistics. Boston, MA, USA: Pearson Addison Wesley; 2008

[6] Hanif M, Ahmed M, Ahmed AM. Biostatistics for health students with manual on software applications. Islamic Society of Statistical Sciences. 2006

[7] Rowe P. Essential Statistics for the Pharmaceutical Sciences. John Wiley & Sons; 2015

[8] Blbas HT, Aziz KF, Nejad SH, Barzinjy AA. Phenomenon of depression and anxiety related to precautions for prevention among population during the outbreak of COVID-19 in Kurdistan region of Iraq: Based on questionnaire survey. Journal of Public Health. 2020; **10**:1-5