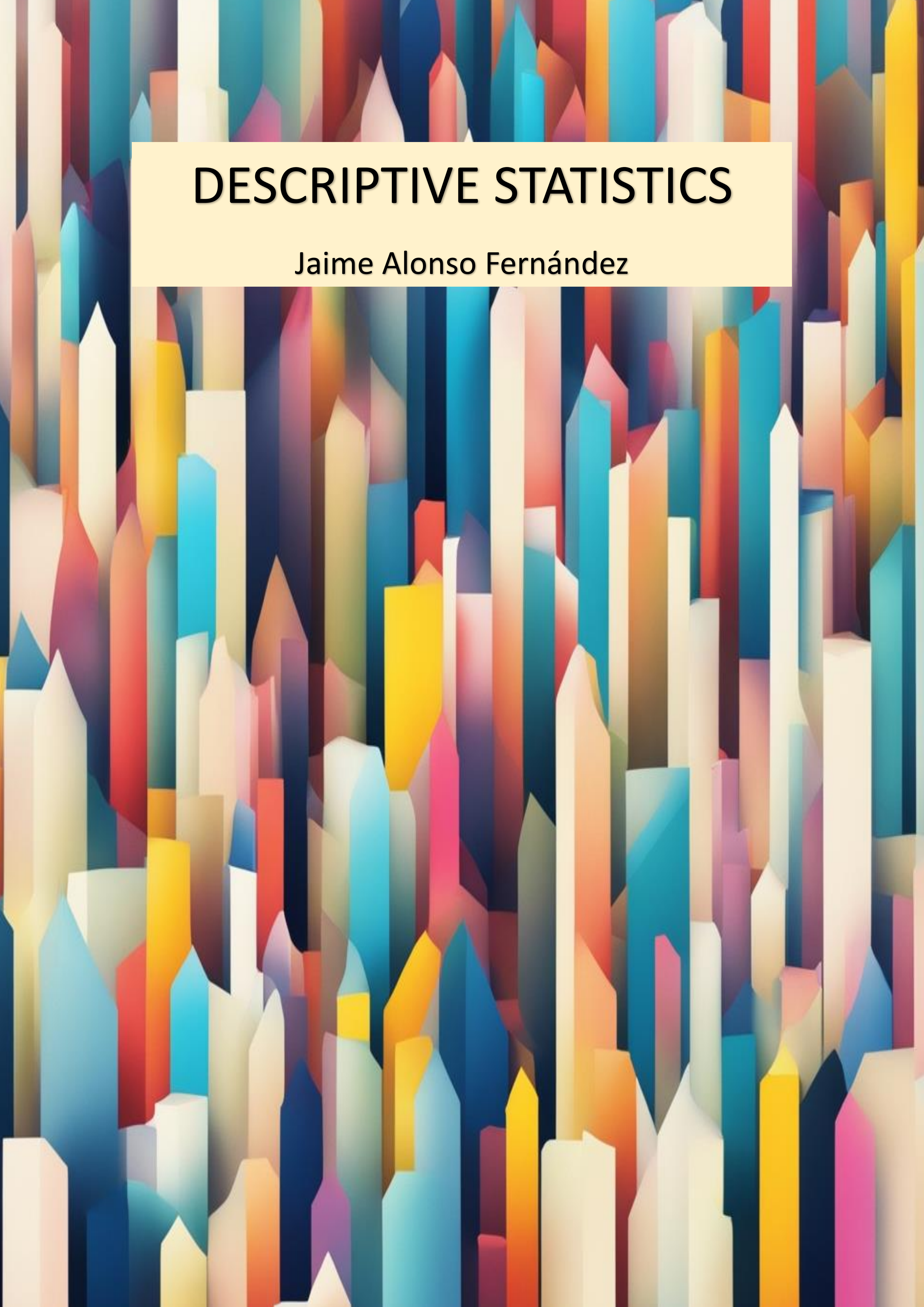
DESCRIPTIVE STATISTICS

Jaime Alonso Fernández



Contents

1. Definition:

Descriptive Statistics are the part of statistics devoted to the organization, representation, and summary of data.

The tool to perform this study will depend on the data to study.

The set of elements to study will be called population, the selected individuals are called subset and on them a variable is observed.

For example, if we want to study the results of some local elections, the population will be the people of the city with the ability to vote, the sample would be the people that are asked, since we will not ask to all the citizens. The variable to study would be which party they voted.

These variables to study, can be of two primary types, both of them divided into two types. In next section will talk about them.

2. Types of Variables

In statistics, different variables to study cannot be studied in the same way. For example, given the number of students with a mark over a 5 and the total number of students, we can easily get the percentage of students that have passed. But only with this data, we cannot know how many students have a mark over a 7 or how many students got an A. (Specially, if the value of the A is not defined). So to differentiate the types of variables, we will use the following classification.

2.1 Quantitative variables:

This, refer to all those variables that are represented numerically. One example of this could be the mark of students in a class, as seen in the previous example.

Depending on how well defined these numerical values are, we divide them into two groups:

2.1.1 Discrete:

This will apply to all the defined number (integers) , for example 10 kids or 5 apples.

2.1.2 Continuous:

This will apply to all the non-defined numbers (decimals), for example the results of an alcohol test.

2.2 Qualitative variables:

This refer to all those variables that are not represented numerically. An example could be the mark system between A-F or the results of an opinion poll from Very Bad to Very Good.

This can also be divided into 2 groups, being these:

2.2.1 Categorical:

Those adjectives that do **not** follow a natural order, for example race or gender.

2.2.2 Ordinal:

Those variables that **do** follow a natural order, for example a grading system from A-F.

3. Frequency tables:

Said to be a table containing different values of the database in order such (from left to right):

- Dataset (**Xi**): Data to study

- Absolute Frequencies (**ni**): Number of times we have category Xi in the dataset, holding that **n1 + n2 + …. + nk = n** where n is the sample size.

- Relative frequencies (**fi**): Proportion of times we have category Xi in the dataset, holding that **fi = ni/n** and **f1 + f2 + …. + fk = 1**.

- Cumulative Absolute Frequency (**Ni**): number of times we observe Xi in the dataset or something smaller, holding **Nk = n.**

- Cumulative Relative Frequency (**Fi**): proportion of times we observe Xi or something smaller, holding **Fi = Ni/n** and **Fk = 1.**

An example could be a study where we what to know how good the results of an exam where. We know that the marks go from 1 to 4 and that they repeated as following:

(1- 5 times, 2- 2 times, 3- 3 times, 4- 6 times)

Xi	**ni**	**fi**	**Ni**	**Fi**
1	5	5/16	5	5/16
2	2	2/16	7	7/16
3	3	3/16	10	10/16
4	6	6/16	16	16/16

4. Graphical Representations

The numerical data may sometimes be hard to process and imagine the proportions. To facilitate said task, we will make use of graphical representations, showing the proportions of the data against all the other.

4.1 Types of Graphical Representations

The type of graphical representation will depend on the type of variable.

For **qualitative** or **quantitative discrete** variables, we shall use:

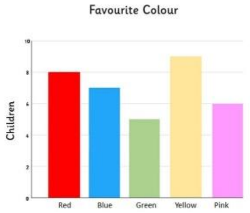
- Bar charts
- Pie Charts

On the other hand, for **quantitative continuous** we shall use:

- Histogram
- Box Plots

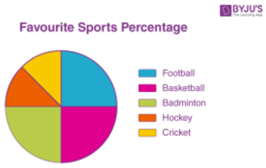
4.1.1 Bar Charts:

It is an X-Y representation, with the **values** laying on the **X axis** and the **frequencies** will lay on the Y axis. The **height** of each bar will be proportional to the frequency.



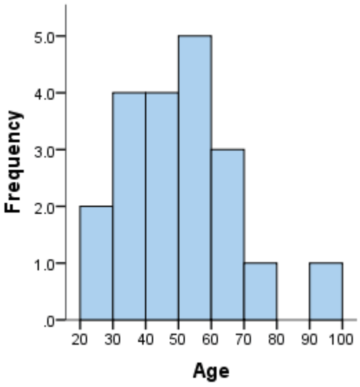
4.1.2 Pie Charts:

We divide a circle in as many areas as **different values** in the dataset, so that the are associate to each value is **proportional** to the **frequency** of the value. For this the frequency of the sector associated with **Xi** should be 360º.



4.1.3 Histograms:

Similarly to a Bar Chart, the **Histogram** will follow an X-Y representation. The **X axis** will represent values **grouped in intervals** and on each interval we will draw a rectangle whose **height** is related to the **frequency**. There are three types:



- **Frequencies** (the **height** is the **number** of data)
- **Percentages** (The **height** is the **proportion** of data within the interval) an
- **Densities** (The **area** is the **proportion** within the interval)

4.1.4 Box Plots:

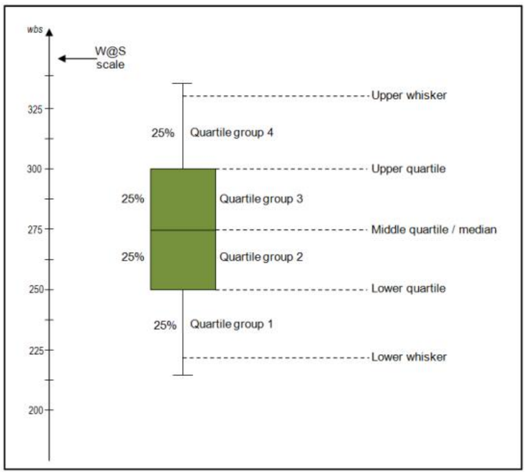
Graphical representation for **quantitative continuous** variables, depicting a few descriptive measures. We have a box from **Q1 to Q3 (quantiles)**. This box will show the value of the median inside.

The **whiskers** go until the smallest and the largest data that **are not outliers.** An observation is an outlier if it is outside the interval:

$$[Q1 - 1.5 * IQR, Q3 + 1.5 * IQR]$$

This measure makes use of **IQR** which will be seen a bit onward.

Outliers will be represented with circles.



5. Descriptive Measures

They are the values that provide some kind of information about the data. There are 3 types:

1. Measures of **central tendency** -> They give the **central value** of the dataset.
2. Measures of **position** -> We **order** the data and give the ones in particular **position**.
3. Measures of **dispersion** -> They tell us if the **values** in the dataset are **similar** or not.

5.1 Measures of Central Tendency

5.1.1 The Mean or Average:

It is the **arithmetic** mean of the data in our dataset. That is for a data [x1, x2, …, xn] , the mean value will be:

$$\bar{X} = \frac{x1 + x2 + \cdots + \ xn}{n}$$

The mean will follow the following characteristics:

1. The mean value will be lower than the maximum and higher than the minimum value, **min <= X <= max**.
2. It can only be computed with **quantitative** variables.
3. If we add a constant K to all data, the new mean will be the same as before plus K.

$$\bar{\mathrm{X}} + \mathrm{K} = \frac{(x1 + K) + (x2 + K) + \cdots + (xn + K)}{n}$$

4. If we multiply all data by K, the new mean will be the same as before times K.
5. It is **very** sensitive to **outliers**[1].

An example of all of this properties, would be:

Given a sample space X with a mean value of 3, if we added the value 1 to all of the elements in the sample space, the mean value would then be 4.

[1] An outlier is a value that is bigger or smaller than the average in a notable way.

Same happens if for example we multiplied all the values of X times 2, we would get a new mean that would be 3*2 = 6.

If to this sample space X we appended a new value, call it xn, such that xn is really big compared to the rest of the values inside X, the mean would be greatly increased, thus the mean is really sensitive to outliers.

5.1.2 The Median:

We order the dataset and pick the value in the middle. If the ordered dataset is:

$$X(1) \leq \mathrm{X}(2) \leq \cdots \leq X(n-1) \leq X(n)$$

Where X(1) is the minimum value in the data set and X(n) is the maximum value, the median will be:

For an odd number of values:

$$\tilde{X} = X(\frac{n+1}{2})$$

For an even number of values in the dataset:

$$\tilde{\mathrm{X}} = \frac{x\left(\frac{n}{2}\right) + x(\left(\frac{n}{2}\right) + 1)}{2}$$

Remember that **X(j)** stands for the value on the **ordered** data set list X(1)….X(n) with the **position j**. That is for the odd case, the **position** will be the one obtained in the formula meanwhile the **value** will be checked in the **ordered value list**.

The median follows the following characteristics:

1. The median will always be greater than the lowest value and smaller than the greatest.

2. It can only be computed with **quantitative** variables.

3. If we a add a constant K to all the data, the new median will be the one before plus K.

4. If we a multiply a constant K to all the data, the new median will be the one before times K.

5. It is **less sensitive to outliers** than the mean.

These properties will apply in the same way as in the mean with exception of the outlier one, since when a value is appended to the list of elements, even though it is really big compared to the rest, the median will remain practically equal if not the same.

Computing the median could be a bit tricky. For that is typical to use a **table of frequencies** where the cumulative relative frequency will help us.

In the case there does not exist a value that is exactly 50%, we would get the next one, in the other case we would get the value that is exactly 50% and the next one and compute the mean.

5.1.3 Mode:

The mode is said to be the most repeated value in the sample space.

5.2 Measures of position: Quantiles

We will fix a position α in [0,1].

The **quantile** is the value with a proportion α of data below in the ordered sample (and 1-α above).

In other words, the value that we shall get after computing the quantile is that value in an ordered list that has a percentage of data (α) below it. (and so, a percentage of 1-α above).

For a value α = 0.5, it will be the same that the as the **median**.

For a value α = 0.25, we will have the **first quartile(C1).**

For a value α = 0.75, we will have the **third quartile(C3)**.

These values are obtained through a **table of frequencies**, by checking the column of the relative frequencies **(Fi).** In the case there is not an exact value for that percentage, you shall get the closes value **above** it.

Ex. Wanting to compute the quantile α = 0.3 (That is 30%) of a set of values, we shall check the frequency table. We see that for the first row we have Fi = 0.2 and for the second row we have Fi = 0.4.We shall the get the value corresponding to the second row. This is because we assure that the **number** were Fi=0.3 will be in a repetition of that value.

5.3 Measures of dispersion

Measures of dispersion are the ones in charge of measuring the similarity among the values. We have 2 types:

5.3.1 Range:

It is the **difference** between the **maximum** and **minimum** values. This makes it really **sensitive** to **outliers**.

$$r = \max(x) - \min(x)$$

5.3.2 Interquartile Range (IQR):

It is the difference between the **third quartile** and the **first quartile**. This makes it **not sensitive to outliers.**

$$IQR = Q3 - Q1$$

5.3.3 Variance:

The **variance** measures the **distance** from each value to the mean and it is usually used to determine the volatility of the sample.

$$S^2 = \frac{\Sigma(xi - \overline{X})^2}{n - 1}$$

- The **variance** will be **0** if and only if **all the data coincide**.

- If we **add** a **constant**, the variance will **not change**.

- If we **multiply** all data times a **constant**, the **variance** will be **multiplied by the square** of said **constant**.

For example, in a data set such [2, 2, 3, 4, 4, 6, 7], the mean value would be 4. The variance will then be:

$$S^2 = \frac{(2-4)^2 + (2-4)^2 + (3-4)^2 + (4-4)^2 + (4-4)^2 + (5-4)^2 + (6-4)^2 + (7-4)^2}{7-1} = 3{,}83$$

So that means that the variance is 3,83.

5.3.4 Standard Deviation:

The **standard deviation** measures the variation of a **random variable** expected about **its mean**. It **differs** from the standard deviation in the way that it does not measure the distance from each individual point but the **'mean' distance to the mean. Low** standard deviation **means** that **all the values are close to the mean**.

$$S = \sqrt{S^2}$$

- If we multiply the data times a constant, the standard deviation will be multiplied by said constant.

- The rest of the properties will be the same as the ones in the variance.

Recalling the previous example, the Standard Deviation will then be:

$$S = \sqrt{3{,}83} = 1{,}96$$

Meaning then that the most common values are between 2,04 and 5,96.

5.3.5 Coefficient of variation:

The coefficient of variation is said to be a standardized measure that represent the ratio Standard deviation – Mean. The objective of this coefficient of variation is to tell how accurate is the mean value (or how representative).

$$cv = \frac{S}{\overline{X}}$$

An example of this, would be for two different samples, one with a cv = 200 and the other with a cv = 50, we can assume that the mean for the sample 2 will give us more significant results than the one of sample 1.