Chapter 5

Descriptive Statistics

Statistics is a branch of Mathematics that deals with the collection, analysis, display and interpretation of numerical data. It consists of two main areas:
Descriptive Statistics includes the collection, presentation and description of numerical data. It is what most people think of when they hear the word "Statistics".
Inferential Statistics consists of the techniques of interpretation, of modeling the results from descriptive Statistics and then using them to make inferences.

1 Analysis and Display of Data

1.1 Basic Concepts

A **population** is a set of individuals, objects, items or measurements whose properties are to be analyzed.
In order to form a population, a set must have a common feature. The population of interest must be carefully defined and is considered so when its membership list is specified.

A subset of the population is called a **sample**, or a **selection**. A sample must be random (each element of the population must have the same chance of being chosen) and representative for the population it was drawn from (the structure of the sample must be similar to the structure of the population).

A **characteristic** or **variable** is a certain feature of interest of the elements of a population or a sample, that is about to be analyzed statistically. Characteristics can be *quantitative* (numerical) or *qualitative* (a certain trait).

From the probabilistic point of view, a numerical characteristic is a random variable. Further, numerical variables can be *discrete* (if they can be counted) or *continuous* (if they can be measured). A numerical characteristic is called a **parameter**, if it refers to an entire population and a **statistic**, if it refers just to a sample.

The outcomes of an experiment yield a set of **data**, i.e. the values that a variable takes for all the elements of a population or a sample.

1.2 Data Collection, Sampling

An important first step in any statistical analysis is the **sampling technique**, i.e. the collection of methods and procedures used to gather data. There are several ways of collecting data: If every element of a population is selected, then a **census** is compiled. However, this technique is hardly ever used these days, because it can be expensive, time consuming or just plain impossible. Instead, only a **sample** is selected, which is analyzed and based on the findings, inferences are made about the entire population.

A sample is chosen based on a **sampling design**, the process used to collect sample data. If elements are chosen on the basis of being "typical", then we have a **judgment sample**, whereas if they are selected based on probability rules, we have a **probability sample**. Statistical inference requires probability samples. The most familiar probability sample is a **random sample**, in which each possible sample of a certain size has the same chance of being selected and every element in the population has an equal probability of being chosen.

Other types of samples may be considered:

- *systematic* sample
- *stratified* sample
- *quota* sample
- *cluster* sample

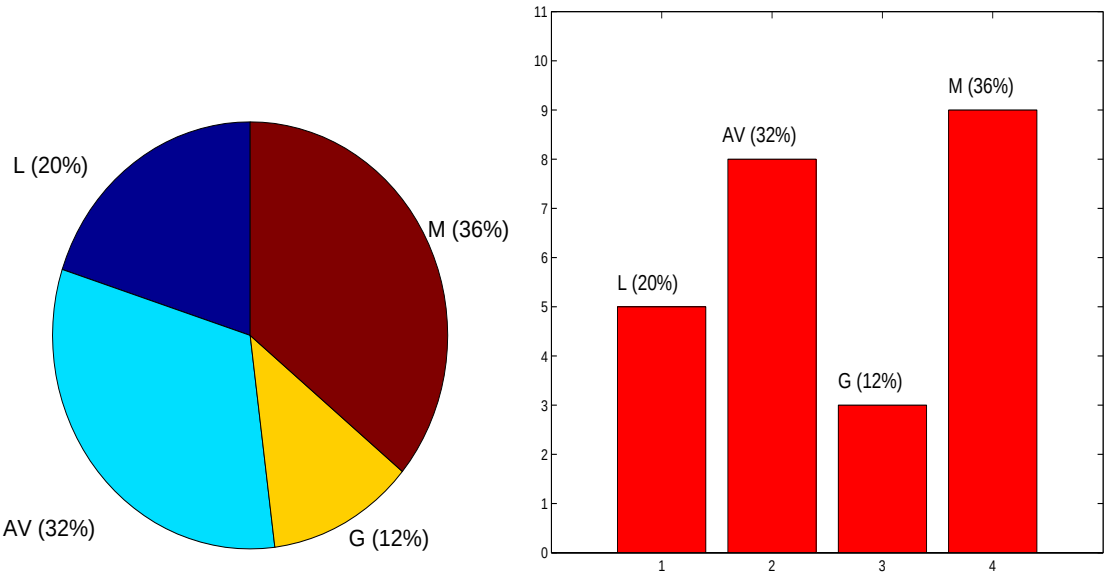
Throughout the remaining chapters, we will only consider random samples.

1.3 Graphical Display of Data, Frequency Distribution Tables, Histograms

"A picture is worth a thousand words!"
Once the sample data is collected, it must be represented in a relevant, "easy to read" way, one that hopefully reveals important features, patterns of behavior, connections, etc.
Circle graphs ("pie" charts) and **bar graphs** are popular ways of displaying data, that use the proportions of each type of data and represent them as percentages.

Example 1.1. Suppose that a software company is having 25 items on sale, 5 of which are learning programs (L), 8 are antivirus programs (AV), 3 are games (G) and the rest (9) are miscellaneous (M).

The pie chart and the bar graph are shown in Figure 1.



(a) Pie Chart (b) Bar Graph

Fig. 1: Example 1.1

Frequency Distribution Tables

Once collected, the raw data must be "organized" in a relevant and meaningful manner. One way to do that is to write it in a **frequency distribution table**, which contains the values $x_i, i = \overline{1,k}$, sorted in increasing order, together with their **(absolute) frequencies**, $f_i, i = \overline{1,k}$, i.e. the number of times each value occurs in the sample data, as seen in Table 1.

Value	Frequency
x_1	f_1
x_2	f_2
$\vdots$	$\vdots$
x_k	f_k

Table 1: Frequency Distribution Table

If needed, the table can also contain the **relative frequencies**

$$rf_i = \frac{f_i}{N},\ \forall i = \overline{1,k},$$

usually expressed as percentages, the **cumulative frequencies**

$$F_i = \sum_{j=1}^{i} f_j,\ \forall i = \overline{1,k},$$

or **relative cumulative frequencies**

$$rF_i = \frac{1}{N}\sum_{j=1}^{i} f_j,\ \forall i = \overline{1,k},$$

where $N = \sum_{i=1}^{k} f_i$ is the sample size.

However, when the data volume is large and the values are nonrepetitive, the frequency distribution is not of much help. Every value is listed with a frequency of 1. In this case, it is better to *group* the data into *classes* and construct a **grouped frequency distribution table**. So, first we decide on a reasonable number of classes n, small enough to make our work with the data easier, but still large enough to not lose the relevance of the data. Then

for each class $i = \overline{1, n}$, we have
$-$ the **class limits** c_{i-1}, c_i,
$-$ the **class mark** $x_i = \dfrac{c_{i-1} + c_i}{2}$, the midpoint of the interval, as an identifier for the class,
$-$ the **class width (length)** $l_i = c_i - c_{i-1}$,
$-$ the **class frequency** f_i, the sum of the frequencies of all observations x in that class.
Notice that we used the same notation x_i for primary data and for class marks. This is by choice, since in the case of grouped data, the class mark plays the role of a "representative" for that class and the class frequency is taken as being the frequency of that one value. The double notation should not cause confusion throughout the text, since N is the sample size, so $x_1, \ldots, x_N$ denotes the primary data, while n is the number of classes and thus,

$$\begin{pmatrix} x_i \\ f_i \end{pmatrix}_{i=\overline{1,n}}$$

denotes the grouped frequency distribution of the data.

The grouped frequency distribution table will look similar to the one in Table 1, only it will contain classes instead of individual values, each with their corresponding features.

Remark 1.2.
1. Relative or cumulative frequencies can also be computed for grouped data, as well, using the same formulas as for ungrouped data.
2. In general, the classes are taken to be of the same length l.
3. When all classes have the same length, the number of classes, n, and the class length l determine each other (if one is known, so is the other). In this case, there are two customary procedures (empirical formulas) of determining the number of classes:
One is a formula for n, known as *Sturges' rule*

$$n = 1 + \frac{10}{3}\log_{10} N, \tag{1}$$

where N is the sample size. Then it follows that $l = \dfrac{x_{\max} - x_{\min}}{n}$.

The other is a formula for the class width

$$l = \frac{8}{100}\left(x_{\max} - x_{\min}\right). \tag{2}$$

Then $n = \dfrac{x_{\max} - x_{\min}}{l}$.
Once we determined n and l, we have $c_i = x_{\min} + i \cdot l,\ i = \overline{0, n}$.

Histograms and Frequency Polygons
When data is grouped into classes, the best way to visualize the frequency distribution is by constructing a **histogram** (hist). A histogram is a type of bar graph, where classes are represented by rectangles whose bases are the class lengths and whose heights are chosen so that the areas of the rectangles are proportional to the class frequencies. If the classes have all the same length, then the heights will be proportional to the class frequencies. If relative frequencies are considered (so the proportionality factor is N, the total number of observations), then the total areas of all rectangles will be equal to 1. For a large volume of data grouped into a reasonably large number of classes, the histogram gives a rough approximation of the density function of the population from which the sample data was drawn.

An alternative in that sense (the sense of roughly approximating the shape of the density function) to histograms are **frequency polygons**, obtained by joining the points with coordinates $(x_i, f_i),\ i = \overline{1, n}$ ($x-$coordinates are the class marks and $y-$coordinates are the class frequencies).

Example 1.3. The following represents the grades distribution in a Probability and Statistics exam, for one section of 2^{nd} year students:

$$\begin{array}{cccccccccccccccccc} 7 & 8 & 10 & 5 & 4 & 5 & 5 & 6 & 5 & 8 & 9 & 9 & 1 & 4 & 5 & 5 & 7 & 10 \\ 5 & 9 & 2 & 2 & 10 & 10 & 8 & 3 & 8 & 7 & 5 & 6 & 7 & 8 & 9 & 9 & 9 & 4. \end{array}$$

Let us analyze these data. First, we sort them in increasing order:

$$\begin{array}{cccccccccccccccccc} 1 & 2 & 2 & 3 & 4 & 4 & 4 & 5 & 5 & 5 & 5 & 5 & 5 & 5 & 5 & 6 & 6 & 7 \\ 7 & 7 & 7 & 8 & 8 & 8 & 8 & 8 & 9 & 9 & 9 & 9 & 9 & 9 & 10 & 10 & 10 & 10 \end{array}$$

There are $N = 36$ observations, with $x_{\min} = 1$ and $x_{\max} = 10$.
Since the sample size is not too large and there are repetitions, we can construct the ungrouped frequency distribution table:

Value	Frequency
1	1
2	2
3	1
4	3
5	8
6	2
7	4
8	5
9	6
10	4

Table 2: Frequency Distribution Table

Let us group the data into classes of the same length. With Sturges' rule, we get

$$n = 6.1877 \approx 6, \;\; l = 1.5,$$

while if using formula (2), we have

$$l = 0.72, \;\; n \approx 12.$$

The grouped frequency tables are shown in Tables 3 and 4. We have also included the relative and cumulative frequencies.
Figure 2 shows the corresponding histogram and frequency polygon for grouped data.

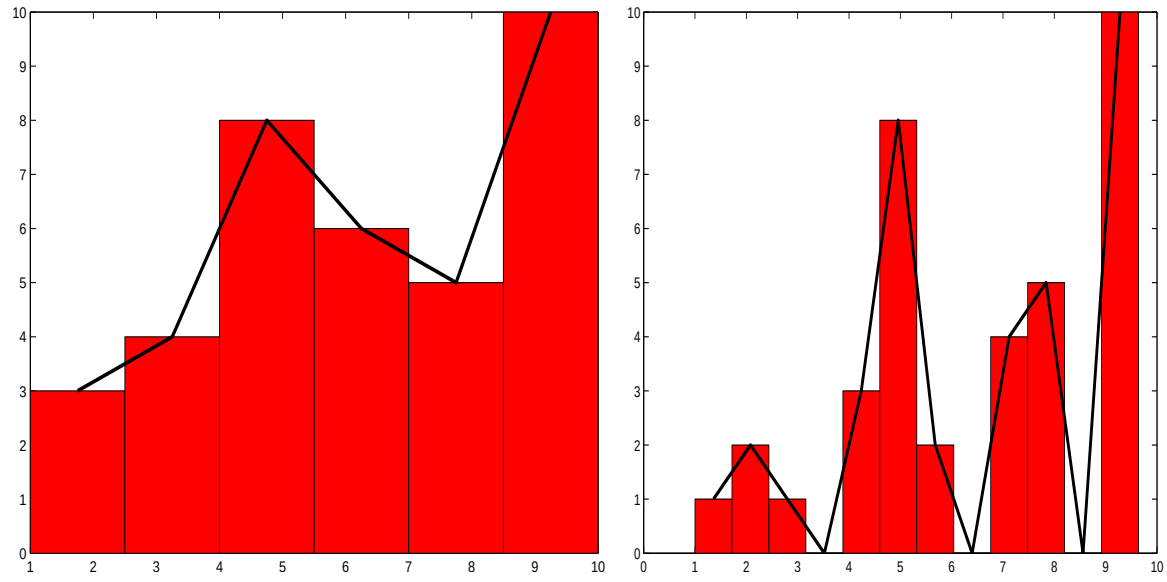
No	Class	Mark	Freq.	C. Freq.	R. Freq.	R. C. Freq.
1	[1.00 , 2.50)	1.75	3	3	0.08%	0.08%
2	[2.50 , 4.00)	3.25	4	7	0.11%	0.19%
3	[4.00 , 5.50)	4.75	8	15	0.22%	0.41%
4	[5.50 , 7.00)	6.25	6	21	0.17%	0.58%
5	[7.00 , 8.50)	7.75	5	26	0.14%	0.72%
6	[8.50 , 10.00]	9.25	10	36	0.28%	1.00%

Table 3: Grouped Frequency Distribution Table With $n = 6$ Classes

Remark 1.4. Due to rounding errors, the length of the last class may be slightly different than the rest of them, even when we group data into classes of the same width.

No	Class	Mark	Freq.	C. Freq.	R. Freq.	R. C. Freq.
1	[1.00 , 1.72)	1.36	1	1	0.03%	0.03%
2	[1.72 , 2.44)	2.08	2	3	0.06%	0.09%
3	[2.44 , 3.16)	2.80	1	4	0.03%	0.12%
4	[3.16 , 3.88)	3.52	0	4	0.00%	0.12%
5	[3.88 , 4.60)	4.24	3	7	0.08%	0.20%
6	[4.60 , 5.32)	4.96	8	15	0.22%	0.42%
7	[5.32 , 6.04)	5.68	2	17	0.06%	0.48%
8	[6.04 , 6.76)	6.40	0	17	0.00%	0.48%
9	[6.76 , 7.48)	7.12	4	21	0.11%	0.59%
10	[7.48 , 8.20)	7.84	5	26	0.14%	0.73%
11	[8.20 , 8.92)	8.56	0	26	0.00%	0.73%
12	[8.92 , 10]	9.46	10	36	0.27%	1.00%

Table 4: Grouped Frequency Distribution Table With $n = 12$ Classes



(a) $n = 6$ (b) $n = 12$

Fig. 2: Histogram and Frequency Polygon

2 Calculative Descriptive Statistics

In the last section, we have considered some graphical methods for getting an idea of the shape of the density function of the population from which the sample data was drawn. Some characteristics, such as symmetry, regularity can be observed from these graphical displays of the data. Next, we consider some statistics that allow us to summarize the data set analytically. It is hoped that these will give us some idea of the values of the parameters that characterize the entire population. We are looking mainly at two types of statistics: *measures of central tendency*, i.e. values that locate the observations with highest frequencies (so, where most of the data values lie) and *measures of variability* that indicate how much the values are spread out.

2.1 Measures of Central Tendency

These are values that tend to locate in some sense the "middle" of a set of data. The term "average" is often associated with these values. Each of the following measures of central tendency can be called the "average" value of a set of data.

Definition 2.1. *The **(arithmetic) mean*** (mean) *of the data $x_1, \ldots, x_N$ is the value*

$$\overline{x}_a = \frac{1}{N} \sum_{i=1}^{N} x_i. \tag{3}$$

For grouped data, $\begin{pmatrix} x_i \\ f_i \end{pmatrix}_{i=\overline{1,n}}$,

$$\overline{x}_a = \frac{1}{N} \sum_{i=1}^{n} f_i x_i.$$

Remark 2.2. Some immediate properties of the arithmetic mean are the following:
1. The sum of all deviations from the mean is equal to 0. Indeed,

$$\sum_{i=1}^{N} (x_i - \overline{x}_a) = \sum_{i=1}^{N} x_i - N\overline{x}_a = 0.$$

2. The mean minimizes the mean square deviation, i.e. for every $a \in \mathbb{R}$,

$$\sum_{i=1}^{N} (x_i - a)^2 \geq \sum_{i=1}^{N} (x_i - \overline{x}_a)^2 .$$

A straightforward computation leads to

$$\begin{aligned}
\sum_{i=1}^{N} (x_i - a)^2 &= \sum_{i=1}^{N} [(x_i - \overline{x}_a) - (a - \overline{x}_a)] \\
&= \sum_{i=1}^{N} (x_i - \overline{x}_a)^2 - 2(a - \overline{x}_a) \sum_{i=1}^{N} (x_i - \overline{x}_a) \\
&+ N \sum_{i=1}^{N} (a - \overline{x}_a)^2 \\
&\geq \sum_{i=1}^{N} (x_i - \overline{x}_a)^2 ,
\end{aligned}$$

since the second term is 0 and the third term is always nonnegative.

Definition 2.3. *The **geometric mean*** (geomean) *of the data $x_1, \ldots, x_N$ is the value*

$$\overline{x}_g = \sqrt[N]{x_1 \ldots x_N}. \tag{4}$$

For grouped data, $\begin{pmatrix} x_i \\ f_i \end{pmatrix}_{i=\overline{1,n}}$,

$$\overline{x}_g = \sqrt[N]{x_1^{f_1} \ldots x_n^{f_n}}.$$

The geometric mean is used in Economics Statistics for price study. One of its distinctive features is that it emphasizes the relative deviations from central tendency, as opposed to the ordinary deviations, emphasized by the arithmetic mean.

Definition 2.4. *The **harmonic mean*** (harmmean) *of the data $x_1, \ldots, x_N$ is the value*

$$\overline{x}_h = \frac{N}{\sum_{i=1}^{N} \frac{1}{x_i}}. \tag{5}$$

For grouped data, $\begin{pmatrix} x_i \\ f_i \end{pmatrix}_{i=\overline{1,n}}$,

$$\overline{x}_h = \frac{N}{\sum_{i=1}^{n} \frac{f_i}{x_i}}.$$

The harmonic mean has applications in Economics Statistics in the study of time norms.

Remark 2.5.
1. For any set of data $x_1, \ldots, x_N$, the well-known *means inequality* holds:

$$\overline{x}_h \leq \overline{x}_g \leq \overline{x}_a,$$

with equality holding if and only if $x_1 = \ldots = x_N$.
2. The most widely used is the algebraic mean. When nothing else is mentioned, we simply say *mean*, instead of *algebraic mean*, and use the simplified notation $\overline{x}$.

Definition 2.6. *The **median*** (median) *is the value x_{me} that divides a set of ordered data X into two equal parts, i.e. the value with the property*

$$P(X < x_{me}) \leq \frac{1}{2} \leq P(X \leq x_{me}). \tag{6}$$

Remark 2.7. The median may or may not be one of the values in the data. If the sorted primary data is

$$x_1 \leq \ldots \leq x_N,$$

then

$$x_{me} = \begin{cases} x_{k+1}, & \text{if } N = 2k+1 \\ \dfrac{x_k + x_{k+1}}{2}, & \text{if } N = 2k \end{cases}.$$

Definition 2.8. *A **mode**, x_{mo}, of a set of data is a most frequent value.*

Remark 2.9.
1. Notice from the wording of the definition that the mode may not be unique. A set of data can have one mode, two modes $-$ *bimodal data*, three

modes − *trimodal data*, or more − *multimodal data*. If every value occurs only once, we say that there is *no mode*.
2. For perfectly symmetric distributions, we have

$$\overline{x} = x_{me} = x_{mo}.$$

This is true, for instance, for the normal distribution. In general,

$$x_{mo} \approx \overline{x} - 3(\overline{x} - x_{me}).$$

2.2 Measures of Variability

Once we have located the "middle" of a set of data, it is important to measure the variability of the data, how much does the data get further away from those middle values. These measures of variation will have small values for closely grouped data (little variation) and larger values for more widely spread out data (large variation).

Consider the primary data $X = \{x_1, \ldots, x_N\}$. The first two measures of variation give a very general idea of the spread in the data values.

Definition 2.10. *The **range*** (range) *of X is the difference*

$$x_{max} - x_{min}.$$

If the values of X are sorted in increasing order, then the range is $x_N - x_1$.

Definition 2.11. *The **mean absolute deviation*** (mad) *of X is the value*

$$MAD = \frac{1}{N} \sum_{i=1}^{N} |x_i - \overline{x}|.$$

Next, following the idea behind the definition of the median, we define values that divide the data into certain percentages.

Definition 2.12. *Let X be a set of data sorted increasingly.*

(1) *The **percentiles*** (prctile) *of X are the values $P_1, P_2, \ldots, P_{99}$ that divide the data into 100 equal parts, i.e. for $k = \overline{1, 99}$, P_k has the property*

$$P(X < P_k) \leq \frac{k}{100}, \ \frac{100 - k}{100} \leq P(X \leq P_k). \tag{7}$$

(2) *The **quartiles** of X are the values*

$$Q_1 = P_{25},\ Q_2 = P_{50} = x_{me}\ \textit{and}\ Q_3 = P_{75}, \tag{8}$$

that divide the data into 4 equal parts.

Remark 2.13. Another important particular case for percentiles are the *deciles*,

$$D_i = P_{10i},\ i = \overline{1,9}.$$

Definition 2.14. *Let X be a set of sorted data with quartiles Q_1, Q_2 and Q_3.*

(1) *The **interquartile range*** ($\boxed{\text{iqr}}$) *is the difference between the third and the first quartile*

$$IQR = Q_3 - Q_1. \tag{9}$$

(2) *The **interquartile deviation** or the **semi interquartile range** is the value*

$$IQD = \frac{IQR}{2} = \frac{Q_3 - Q_1}{2}. \tag{10}$$

(3) *The **interquartile deviation coefficient** or the **relative interquartile deviation** is the value*

$$IQDC = \frac{IQD}{x_{me}} = \frac{Q_3 - Q_1}{2Q_2}. \tag{11}$$

Remark 2.15.
1. The interquartile deviation is an absolute measure of variation and it has an important property: the range $x_{me} \pm IQD$ contains approximately 50% of the data.
2. The interquartile deviation coefficient $IQDC$ varies between -1 and 1, taking values close to 0 for symmetrical distributions, with little variation and values close to ± 1 for skewed data with large variation.

The interquartile range is also involved in another important aspect of statistical analysis, namely the detection of outliers. An *outlier* , as the name suggests, is basically an atypical value, "far away" from the rest of the data, that does not seem to belong to the distribution of the rest of the values in

the data set. For example, in set of data where all values are between 0 and 1, a value of 1000 would surely seem out of place. Outliers can arise for two reasons: either they are legitimate observations whose values are simply unusually large or unusually small, compared to the rest of the values in the data set, or they are the result of an error in measurement, of poor experimental techniques, or of mistakes in recording or entering the data. Whichever the reason, they can adversely affect some values of the measures of central tendency and of variation, thus leading to erroneous inferential results. Once the presence of such outliers is detected, it is suggested that sample statistics be computed both with and without the outliers. Thus the problem of detecting and locating an outlier is an important part of any statistical data analysis process. For instance, one simple procedure would be to consider an outlier any value that is more than 2.5 standard deviations away from the mean, and an extreme outlier a value more than 3 standard deviations away from the mean. This procedure is justified by the "3σ rule" and would work well for unimodal and symmetrical distributions. A more general approach, that works for skewed data, is to consider an outsider any observation that is outside the range

$$\left[Q_1 - \frac{3}{2}IQR,\ Q_3 + \frac{3}{2}IQR\right] = \left[Q_1 - 3IQD,\ Q_3 + 3IQD\right].$$

Example 2.16. Consider the following set of data

$$\begin{array}{rrrrr} 0.5973 & 0.3624 & 0.8304 & 1.7347 & 1.2499 \\ 0.1104 & 0.8082 & 0.6039 & 0.3046 & 0.6183 \\ 0.0065 & 0.8748 & 1.3528 & 1.6458 & 1.5117 \\ 0.3253 & -2.0000 & -1.3000 & 1.7500 & 3.8500 \end{array}$$

We sort them in increasing order:

$$\begin{array}{rrrrr} -2.0000 & -1.3000 & 0.0065 & 0.1104 & 0.3046 \\ 0.3253 & 0.3624 & 0.5973 & 0.6039 & 0.6183 \\ 0.8082 & 0.8304 & 0.8748 & 1.2499 & 1.3528 \\ 1.5117 & 1.6458 & 1.7347 & 1.7500 & 3.8500 \end{array}$$

We have:

$$\begin{aligned} Q1 &= 0.3150, \\ Q2 &= 0.7133, \\ Q3 &= 1.4323, \end{aligned}$$

$$Q_1 - \frac{3}{2}IQR = -1.3610,$$

$$Q_3 + \frac{3}{2}IQR = 3.1082.$$

The data (boxplot) is displayed graphically in Figure 3.

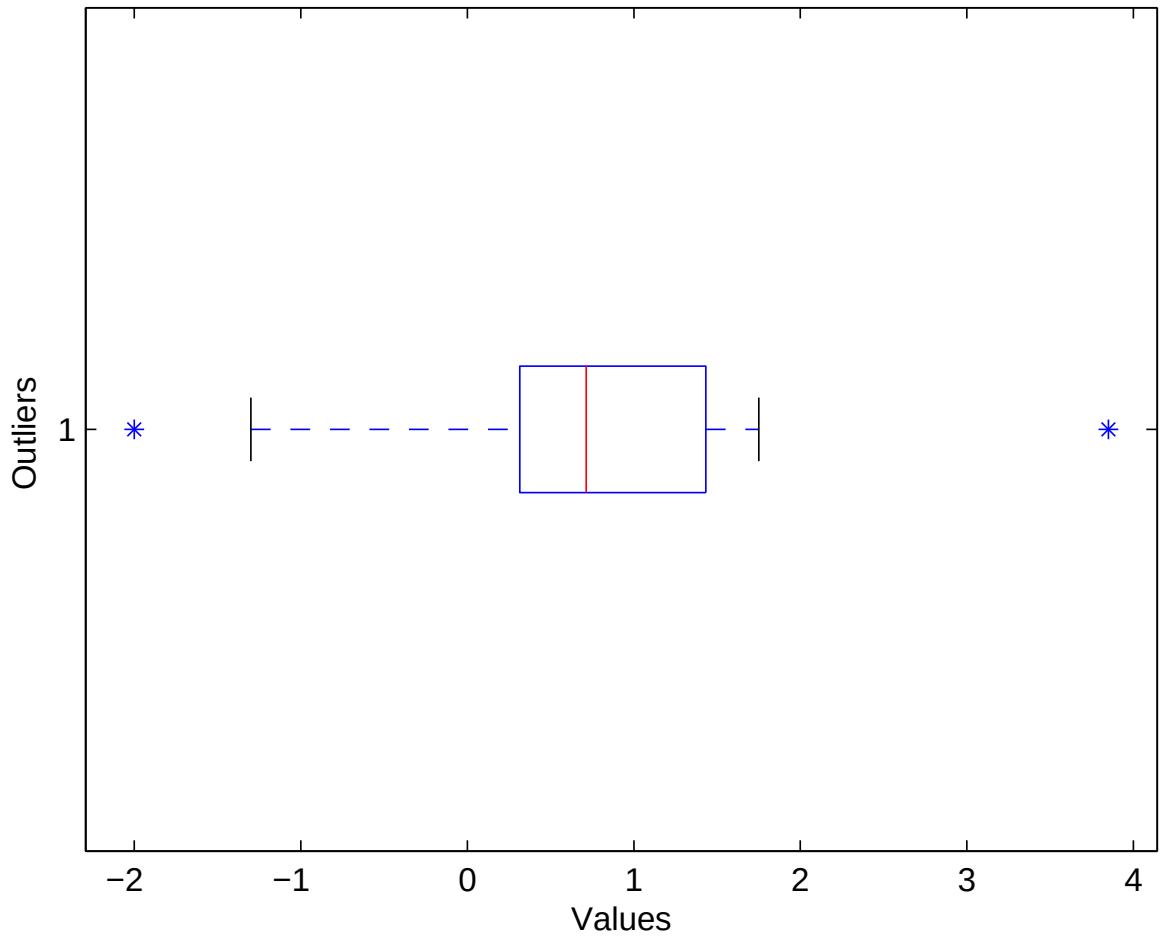


Fig. 3: Quartiles, Interquartile Range, Outliers

Definition 2.17.

(1) *The **moment of order k** is the value*

$$\overline{\nu}_k = \frac{1}{N}\sum_{i=1}^{N} x_i^k, \;\; \overline{\nu}_k = \frac{1}{N}\sum_{i=1}^{n} f_i x_i^k, \tag{12}$$

for primary and for grouped data, respectively.

(2) *The **central moment of order k*** (moment) *is the value*

$$\overline{\mu}_k = \frac{1}{N}\sum_{i=1}^{N} (x_i - \overline{x})^k, \;\; \overline{\mu}_k = \frac{1}{N}\sum_{i=1}^{n} f_i (x_i - \overline{x})^k \tag{13}$$

for primary and for grouped data, respectively.

(3) *The **variance*** (var) *is the value*

$$\overline{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N} (x_i - \overline{x})^2, \;\; \overline{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{n} f_i (x_i - \overline{x})^2 \tag{14}$$

*for primary and for grouped data, respectively. The quantity $\overline{\sigma} = \sqrt{\overline{\sigma}^2}$ is the **standard deviation*** (std).

Remark 2.18.

1. We will see later that when the data represents a sample (not the entire population), a better formula would be

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N} (x_i - \overline{x})^2, \;\; s^2 = \frac{1}{N-1}\sum_{i=1}^{n} f_i (x_i - \overline{x})^2, \tag{15}$$

for the sample variance for primary or grouped data. The reason for that will have to do with the "bias" and will be explained later on in the next chapter. For now, we will just agree to use (14) to compute the variance of a set of data that represents a population and (15) for the variance of a sample.

2. A more efficient computational formula for the variance is

$$\overline{\sigma}^2 = \frac{1}{N}\left(\sum_{i=1}^{N} x_i^2 - \frac{1}{N}\left(\sum_{i=1}^{N} x_i\right)^2\right), \tag{16}$$

which follows straight from the definition.

Definition 2.19. *The **coefficient of variation** is the value*

$$CV = \frac{\overline{\sigma}}{\overline{x}}.$$

Remark 2.20.
1. The coefficient of variation can be expressed as a ratio or as a percentage. It is useful in comparing the degrees of variation of two sets of data, even when their means are different.
2. The coefficient of variation is widely used in Biostatistics and Business Statistics. For example, in the investing world, the coefficient of variation helps brokers determine how much volatility (risk) they are assuming in comparison to the amount of return they can expect from a certain investment. The lower the value of the CV, the better the risk-return tradeoff.

3 Correlation and Regression

So far we have been discussing a number of descriptive techniques for describing one variable only. However, a very important part of statistics is describing the association between two (or more) variables, whether or not they are independent, and if they are not, what is the nature of their dependence. One of the most fundamental concepts in statistical research is the concept of correlation.
Correlation is a measure of the relationship between one dependent variable and one or more independent variables. If two variables are correlated, this means that one can use information about one variable to predict the values of the other variable. **Regression** is then the method or statistical procedure that is used to establish that relationship.

3.1 Correlation, Curves of Regression

We will restrict our discussion to the case of two characteristics, X and Y. If X and Y have the same length, we can get a first idea of the relationship between the two, by plotting them in a **scattergram**, or **scatterplot**, , which is a plot of the points with coordinates $(x_i, y_i)_{i=\overline{1,k}},\ x_i \in X,\ y_i \in Y,\ i = \overline{1,k}$.

We group the N primary data into mn classes and denote by (x_i, y_j) the class mark and by f_{ij} the absolute frequency of the class (i, j), $i = \overline{1, m}$, $j = \overline{1, n}$. Then we represent the two-dimensional characteristic (X, Y) in a *correlation table*, or *contingency table*, as shown below.

$X \setminus Y$	y_1	$\dots$	y_j	$\dots$	y_n	
x_1	f_{11}	$\dots$	f_{1j}	$\dots$	f_{1n}	$f_{1.}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_i	f_{i1}	$\dots$	f_{ij}	$\dots$	f_{in}	$f_{i.}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_m	f_{m1}	$\dots$	f_{mj}	$\dots$	f_{mn}	$f_{m.}$
	$f_{.1}$	$\dots$	$f_{.j}$	$\dots$	$f_{.n}$	$f_{..} = N$

Table 5: Correlation Table

Notice that

$$\sum_{j=1}^{n} f_{ij} = f_{i.}, \ \sum_{i=1}^{m} f_{ij} = f_{.j}, \ \sum_{i=1}^{m} f_{i.} = \sum_{j=1}^{n} f_{.j} = f_{..} = N.$$

Now we can define numerical characteristics associated with (X, Y).

Definition 3.1. *Let (X, Y) be a two-dimensional characteristic whose distribution is given by Table 5 and let $k_1, k_2 \in \mathbb{N}$.*

(1) *The **(initial) moment of order (k_1, k_2)** of (X, Y) is the value*

$$\overline{\nu}_{k_1 k_2} = \frac{1}{N} \sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} x_i^{k_1} y_j^{k_2}. \tag{17}$$

(2) *The **central moment of order (k_1, k_2)** of (X, Y) is the value*

$$\overline{\mu}_{k_1 k_2} = \frac{1}{N} \sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} (x_i - \overline{x})^{k_1} (y_j - \overline{y})^{k_2}, \tag{18}$$

where $\overline{x} = \overline{\nu}_{10} = \dfrac{1}{N} \sum_{i=1}^{m} f_{i.} x_i$ and $\overline{y} = \overline{\nu}_{01} = \dfrac{1}{N} \sum_{j=1}^{n} f_{.j} y_j$ are the means of X and Y, respectively.

Remark 3.2. Just as the means of the two characteristics X and Y can be expressed as moments of (X, Y), so can their variances:

$$\begin{array}{rcccl} \overline{\sigma}_X^2 & = & \overline{\mu}_{20} & = & \overline{\nu}_{20} - \overline{\nu}_{10}^2, \\ \overline{\sigma}_Y^2 & = & \overline{\mu}_{02} & = & \overline{\nu}_{02} - \overline{\nu}_{01}^2. \end{array}$$

Definition 3.3. *Let (X, Y) be a two-dimensional characteristic whose distribution is given by Table 5.*

(1) *The **covariance*** (cov) *of (X, Y) is the value*

$$\operatorname{cov}(X, Y) = \overline{\mu}_{11} = \frac{1}{N} \sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij}(x_i - \overline{x})(y_j - \overline{y}). \tag{19}$$

(2) *The **correlation coefficient*** (corrcoef) *of (X, Y) is the value*

$$\overline{\rho} = \overline{\rho}_{XY} = \frac{\operatorname{cov}(X, Y)}{\sqrt{\overline{\mu}_{20}}\sqrt{\overline{\mu}_{02}}} = \frac{\overline{\mu}_{11}}{\overline{\sigma}_X \overline{\sigma}_Y}. \tag{20}$$

These two notions have been mentioned before, in Chapter 4, for two random variables. They are defined similarly for sets of data and they have the same properties. The covariance gives a rough idea of the relationship between X and Y. As before, if X and Y are independent (so there is no relationship, no correlation between them), then the covariance is 0. If large values of X are associated with large values of Y, then the covariance will have a positive value, if, on the contrary, large values of X are associated with small values of Y, then the covariance will have a negative value. Also, an easier computational formula for the covariance is $\operatorname{cov}(X, Y) = \overline{\nu}_{11} - \overline{x} \cdot \overline{y}$.
The correlation coefficient is then

$$\overline{\rho} = \frac{\overline{\nu}_{11} - \overline{x} \cdot \overline{y}}{\overline{\sigma}_X \overline{\sigma}_Y},$$

as before, satisfies the inequality

$$-1 \leq \overline{\rho} \leq 1 \tag{21}$$

and, by its variation between -1 and 1, its value measures the linear relationship between X and Y. If $\overline{\rho}_{XY} = 1$, there is a *perfect positive correlation* between X and Y, if $\overline{\rho}_{XY} = -1$, there is a *perfect negative correlation*

between X and Y. In both cases, the linearity is "perfect", i.e there exist $a, b \in \mathbb{R},\ a \neq 0$, such that $Y = aX + b$. If $\overline{\rho}_{XY} = 0$, then there is no linear correlation between X and Y, they are said to be *(linearly) uncorrelated*. However, in this case, they may not be independent, some other type of relationship (not linear) may exist between them.

In our task of finding a relationship between X and Y, we may go the following path: knowing the value of one of the characteristics, try to find a probable, an "expected" value for the other. If the two characteristics are related in any way, then there should be a pattern developing, that is the expected value of one of them, conditioned by the other one taking a certain value, should be a function of that value that the other variable assumes. That means we should consider *conditional means*, that were first introduced in Chapter 4.

Definition 3.4. *Let (X, Y) be a two-dimensional characteristic whose distribution is given by Table 5.*

(1) *The **conditional mean** of Y, given $X = x_i$, is the value*

$$\overline{y}_i = \overline{y}(x_i) = \frac{1}{f_{i.}} \sum_{j=1}^{n} f_{ij} y_j,\ i = \overline{1, m}. \tag{22}$$

(2) *The **conditional mean** of X, given $Y = y_j$, is the value*

$$\overline{x}_j = \overline{x}(y_j) = \frac{1}{f_{.j}} \sum_{i=1}^{m} f_{ij} x_i,\ j = \overline{1, n}. \tag{23}$$

Definition 3.5. *Let (X, Y) be a two-dimensional characteristic.*

(1) *The curve $y = f(x)$ formed by the points with coordinates $(x_i, \overline{y}_i)$, $i = \overline{1, m}$, is called the **curve of regression** of Y on X.*

(2) *The curve $x = g(x)$ formed by the points with coordinates $(y_j, \overline{x}_j)$, $j = \overline{1, n}$, is called the **curve of regression** of X on Y.*

Remark 3.6. The curve of regression of a characteristic Y with respect to another characteristic X is then the mean value of Y, $\overline{y}(x)$, given $X = x$. The curve of regression is determined so that it approximates best the scatterplot of (X, Y).

3.2 Least Squares Estimation, Linear Regression

One of the most popular ways of finding curves of regression is the *least squares method*.
Assume the curve of regression of Y on X is of the form

$$y = y(x) = f(x; a_1, \ldots, a_s).$$

We determine the unknown parameters $a_1, \ldots, a_s$ so that the *sum of squares error* (SSE)

$$S = SSE = \sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij}(y_j - y(x_i))^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij}(y_j - f(x_i; a_1, \ldots, a_s))^2$$

is minimum (hence, the name of the method).
We find the point of minimum $(\overline{a}_1, \ldots, \overline{a}_s)$ of S by solving the system

$$\frac{\partial S}{\partial a_k} = 0, \ k = \overline{1, s},$$

i.e.

$$-2 \sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij}(y_j - f(x_i; a_1, \ldots, a_s)) \frac{\partial f(x_i; a_1, \ldots, a_s)}{\partial a_k} = 0, \qquad (24)$$

for every $k = \overline{1, s}$.
Then the equation of the curve of regression of Y on X is

$$y = f\left(x; \overline{a}_1, \ldots, \overline{a}_s\right).$$

Let us consider the case of *linear regression* and find the equation of the *line of regression* of Y on X. We are finding a curve

$$y = ax + b,$$

for which

$$S(a, b) = \sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij}(y_j - ax_i - b)^2$$

is minimum. The system (24) becomes

$$\begin{cases} \left(\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}x_i^2\right) a \;+\; \left(\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}x_i\right) b \;=\; \sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}x_i y_j \\ \left(\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}x_i\right) a \;+\; \left(\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}\right) b \;=\; \sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}y_j \end{cases}$$

and after dividing both equations by N,

$$\begin{cases} \overline{\nu}_{20}a \;+\; \overline{\nu}_{10}b \;=\; \overline{\nu}_{11} \\ \overline{\nu}_{10}a \;+\; \overline{\nu}_{00}b \;=\; \overline{\nu}_{01}. \end{cases}$$

Its solution is

$$\overline{a} \;=\; \frac{\overline{\nu}_{11}-\overline{\nu}_{10}\overline{\nu}_{01}}{\overline{\nu}_{20}-\overline{\nu}_{10}^2} \;=\; \frac{\overline{\nu}_{11}-\overline{x}\cdot\overline{y}}{\overline{\sigma}_X^2} \;=\; \frac{\overline{\nu}_{11}-\overline{x}\cdot\overline{y}}{\overline{\sigma}_X\overline{\sigma}_Y}\cdot\frac{\overline{\sigma}_Y}{\overline{\sigma}_X} \;=\; \overline{\rho}\,\frac{\overline{\sigma}_Y}{\overline{\sigma}_X},$$

$$\overline{b} \;=\; \overline{\nu}_{01}-\overline{\nu}_{10}\overline{a} \;=\; \overline{y}-\overline{a}\cdot\overline{x}.$$

So the equation of the line of regression of Y on X is

$$y-\overline{y}=\overline{\rho}\,\frac{\overline{\sigma}_Y}{\overline{\sigma}_X}\,(x-\overline{x}) \tag{25}$$

and, by analogy, the equation of the line of regression of X on Y is

$$x-\overline{x}=\overline{\rho}\,\frac{\overline{\sigma}_X}{\overline{\sigma}_Y}\,(y-\overline{y})\,. \tag{26}$$

Remark 3.7.
1. The point of intersection of the two lines of regression, $(\overline{x},\overline{y})$, is called the *centroid* of the distribution of the characteristic (X,Y).
2. The slope $\overline{a}_{Y|X}=\overline{\rho}\,\dfrac{\overline{\sigma}_Y}{\overline{\sigma}_X}$ of the line of regression of Y on X is called the *coefficient of regression* of Y on X. Similarly, $\overline{a}_{X|Y}=\overline{\rho}\,\dfrac{\overline{\sigma}_X}{\overline{\sigma}_Y}$ is the coefficient of regression of X on Y and

$$|\,\overline{\rho}\,|=\overline{a}_{Y|X}\,\overline{a}_{X|Y}.$$

3. For the angle α between the two lines of regression, we have

$$\tan\alpha \;=\; \frac{1-\overline{\rho}^2}{\overline{\rho}^2}\cdot\frac{\overline{\sigma}_X\overline{\sigma}_Y}{\overline{\sigma}_X^2+\overline{\sigma}_Y^2}.$$

So, if $\mid \overline{\rho} \mid = 1$, then $\alpha = 0$, i.e. the two lines coincide. If $\mid \overline{\rho} \mid = 0$ (for instance, if X and Y are independent), then $\alpha = \dfrac{\pi}{2}$, i.e. the two lines are perpendicular.

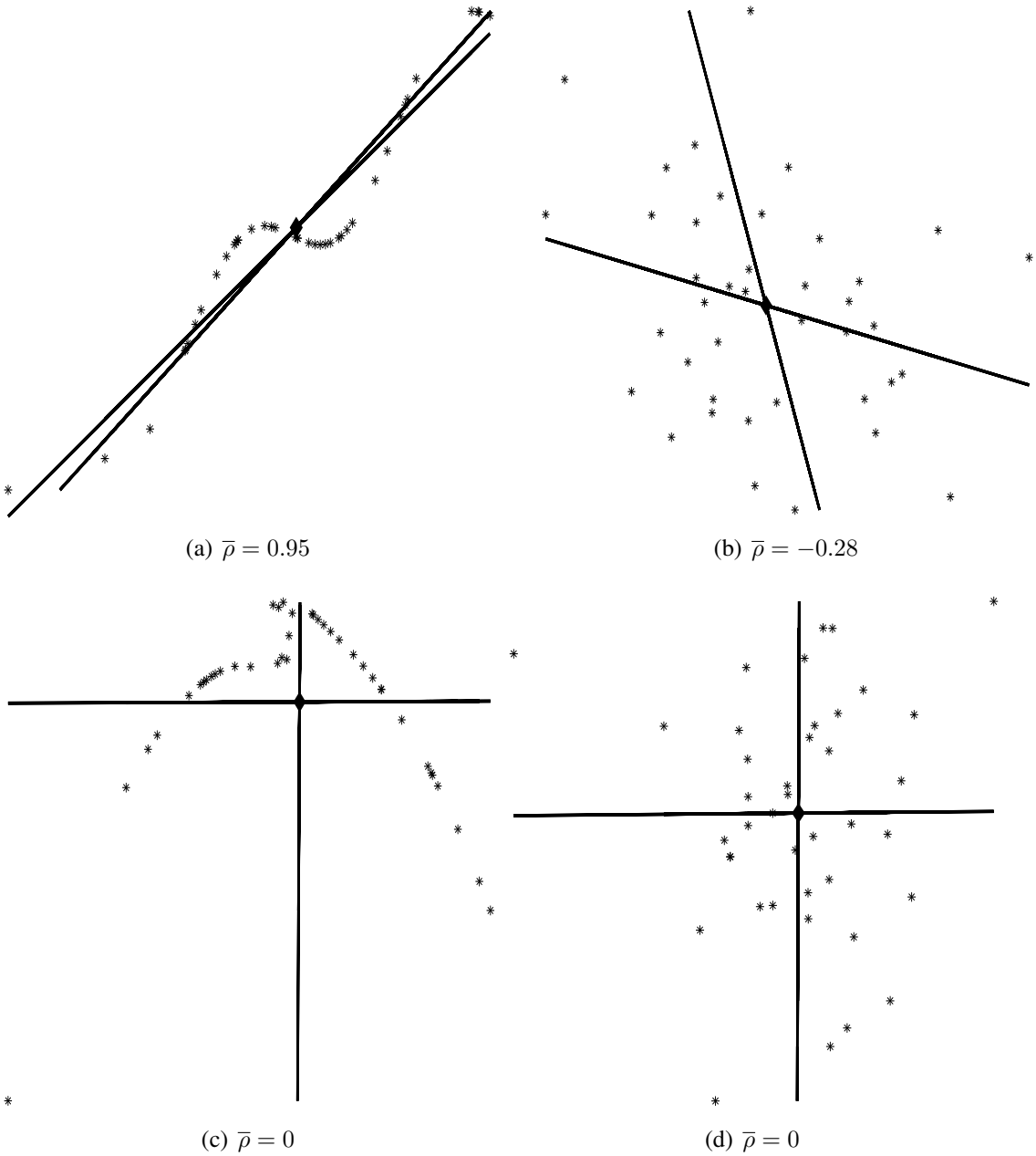
Example 3.8. Let us examine the situations graphed in Figure 4.

- In Figure 4(a) $\overline{\rho} = 0.95$, positive and very close to 1, suggesting a strong positive linear trend. Indeed, most of the points are on or very close to the line of regression of Y on X. The positivity indicates that large values of X are associated with large values of Y. Also, since the correlation coefficient is so close to 1, the two lines of regression almost coincide.

- In Figure 4(b) $\overline{\rho} = -0.28$, negative and fairly small, close to 0. If a relationship exists between X and Y, it does not seem to be linear. In fact, they are very close to being independent, since the points are scattered around the plane, no pattern being visible. The two lines of regression are very distinct and both have negative slopes, suggesting that large values of X are associated with small values of Y.

- In Figure 4(c) $\overline{\rho} = 0$, so the two characteristics are uncorrelated, no linear relationship exists between them. However they are not independent, they were chosen so that $Y = -X^2 + \sin\left(\dfrac{1}{X}\right)$. Notice also, that the two lines of regression are perpendicular.

- Finally, in Figure 4(d) $\overline{\rho} = 0$, again, so no linear relationship exists. In fact the two characteristics are independent, which is suggested by their random scatter inside the plane.



(a) $\overline{\rho} = 0.95$

(b) $\overline{\rho} = -0.28$

(c) $\overline{\rho} = 0$

(d) $\overline{\rho} = 0$

Fig. 4: Scattergram, Lines of Regression and Centroid

Remark 3.9. Other types of curves of regression that are fairly frequently used are
– *exponential* regression $y = ab^x$,
– *logarithmic* regression $y = a \log x + b$,
– *logistic* regression $y = \dfrac{1}{ae^{-x} + b}$,
– *hyperbolic* regression $y = \dfrac{a}{x} + b$.

4 Sample Theory

Suppose we are interested in studying a characteristic (a random variable) X, relative to a population P, of size N. The difficulty or even the impossibility of studying the entire population, as well as the merits of choosing and studying a random sample from which to make inferences about the population of interest, have already been discussed in the previous chapter. Now, we want to give a more rigorous and precise definition of a random sample, in the framework of random variables, one that can then employ probability theory techniques for making inferences.

4.1 Sample Functions

We choose n ($n \leq N$) objects and actually study $X_i,\ i = \overline{1, n}$, the characteristic of interest *for the i^{th} object selected.* Since the n objects were randomly selected, it makes sense that for $i = \overline{1, n}$, X_i is a random variable, one that has the same distribution as X, the characteristic relative to the entire population. Furthermore, these random variables are independent, since the value assumed by one of them has no effect on the values assumed by the others. Once the n objects have been selected, we will have n numerical values available, $x_1, \ldots, x_n$, the observed values of $X_1, \ldots, X_n$.

Definition 4.1. *A **random sample of size n** from the distribution of X, a characteristic relative to a population P, is a collection of n independent random variables $X_1, \ldots, X_n$, having the same distribution as X. The variables $X_1, \ldots, X_n$, are called **sample variables** and their observed values $x_1, \ldots, x_n$, are called **sample data**.*

Remark 4.2. The term *random sample* may refer to the objects selected, to the sample variables, or to the sample data. It is usually clear from the context which meaning is intended. In general, we use capital letters to denote sample variables and corresponding lowercase letters for their values, the sample data.

We are able now to define sample functions, or statistics, in the more precise context of random variables.

Definition 4.3. *A **sample function** or **statistic** is a random variable*

$$Z_n = h_n(X_1, \ldots, X_n),$$

where $h_n : \mathbb{R}^n \to \mathbb{R}$ is a measurable function. The value of the sample function Z_n is $z_n = h_n(x_1, \ldots, x_n)$.

We will revisit now some sample numerical characteristics discussed in the previous sections and define them as sample functions.

Sample Mean

Definition 4.4. *The **sample mean** is the sample function defined by*

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{27}$$

and its value is $\overline{x}_n = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i$.

Now that the sample mean is defined as a random variable, we can discuss its distribution and its numerical characteristics.

Proposition 4.5. *Let X be a characteristic with $E(X) = \mu$ and $V(X) = \sigma^2$. Then*

$$E\left(\overline{X}\right) = \mu \text{ and } V\left(\overline{X}\right) = \frac{\sigma^2}{n}. \tag{28}$$

Moreover, if $X \in N(\mu, \sigma)$, then $\overline{X} \in N\left(\mu, \dfrac{\sigma}{\sqrt{n}}\right)$.

Proof. Since $X_1, \ldots, X_n$ are identically distributed, with the same distribution as X, $E(X_i) = E(X) = \mu$ and $V(X_i) = V(X) = \sigma^2$, $\forall i = \overline{1,n}$. Then, using the properties of expectation, we have

$$E\left(\overline{X}\right) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\, n\mu = \mu.$$

Further, since $X_1, \ldots, X_n$ are also independent, by the properties of variance, it follows that

$$V\left(\overline{X}\right) = V\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} V(X_i) = \frac{1}{n^2}\, n\sigma^2 = \frac{\sigma^2}{n}.$$

The last part follows from the fact that $\overline{X}$ is a linear combination of independent, normally distributed random variables. □

Corollary 4.6. *Let X be a characteristic with $E(X) = \mu$ and $V(X) = \sigma^2$ and for $n \in \mathbb{N}$ let*

$$Z_n = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}.$$

Then the variable Z_n converges in distribution to a standard normal variable, as $n \to \infty$. Moreover, if $X \in N(\mu, \sigma)$, then the statement is true for every $n \in \mathbb{N}$.

Proof. This is a direct consequence of Proposition 4.5 and the CLT. □

Sample Moments and Sample Variance

Definition 4.7. *The statistic*

$$\overline{\nu}_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k \tag{29}$$

*is called the **sample moment of order k** and its value is $\frac{1}{n}\sum_{i=1}^{n} x_i^k$.*

The statistic

$$\overline{\mu}_k = \frac{1}{n}\sum_{i=1}^{n} (X_i - \overline{X})^k \tag{30}$$

is called the ***sample central moment of order k*** *and its value is*
$\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^k$.

Remark 4.8. Just like for theoretical (population) moments, we have

$$\begin{aligned} \overline{\nu}_1 &= \overline{X}, \\ \overline{\mu}_1 &= 0, \\ \overline{\mu}_2 &= \overline{\nu}_2 - \overline{\nu}_1^2. \end{aligned}$$

Next we discuss the distributions and characteristics of these new sample functions.

Proposition 4.9. *Let X be a characteristic with the property that for $k \in \mathbb{N}$, the theoretical moment $\nu_{2k} = \nu_{2k}(X) = E\left(X^{2k}\right)$ exists. Then*

$$E\left(\overline{\nu}_k\right) = \nu_k \text{ and } V\left(\overline{\nu}_k\right) = \frac{1}{n}\left(\nu_{2k} - \nu_k^2\right). \tag{31}$$

Proof. First off, the condition that ν_{2k} exists for X ensures the fact that all theoretical moments of X of order up to k also exist. The rest follows as before. We have

$$E\left(\overline{\nu}_k\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i^k) = \frac{1}{n}\sum_{i=1}^{n} E(X^k) = \frac{1}{n}\, n\nu_k = \nu_k$$

and

$$\begin{aligned} V\left(\overline{\nu}_k\right) &= \frac{1}{n^2}\sum_{i=1}^{n} V(X_i^k) &&= \frac{1}{n^2}\sum_{i=1}^{n} V(X^k) \\ &= \frac{1}{n^2}\, n\left(\nu_{2k} - \nu_k^2\right) &&= \frac{1}{n}\left(\nu_{2k} - \nu_k^2\right). \end{aligned}$$

□

Corollary 4.10. *Let X be a characteristic satisfying the hypothesis of Proposition 4.9 and for $n \in \mathbb{N}$ let*

$$Z_n = \frac{\overline{\nu}_k - \nu_k}{\sqrt{\frac{\nu_{2k} - \nu_k^2}{n}}}.$$

Then Z_n converges in distribution to a standard normal variable.

We only discuss the properties of the sample central moment of order 2.

Proposition 4.11. *Let X be a characteristic with $V(X) = \mu_2 = \sigma^2$ and for which the theoretical moment $\nu_4 = E\left(X^4\right)$ exists. Then*

$$\begin{aligned} E\left(\overline{\mu}_2\right) &= \frac{n-1}{n}\,\sigma^2, \\ V\left(\overline{\mu}_2\right) &= \frac{n-1}{n^3}\left[(n-1)\mu_4 - (n-3)\sigma^4\right], \\ \operatorname{cov}(\overline{X}, \overline{\mu}_2) &= \frac{n-1}{n^2}\,\mu_3. \end{aligned} \tag{32}$$

Proof. We will only prove the first assertion, (32), as it is the most important and oftenly used property of $\overline{\mu}_2$. Using Proposition 4.9, the properties of expectation and the fact that $X_1, \ldots, X_n$ are independent and identically distributed, we have

$$\begin{aligned} E\left(\overline{\mu}_2\right) &= E\left(\overline{\nu}_2\right) - E\left(\overline{\nu}_1^2\right) = \nu_2 - E\left(\left(\frac{1}{n}\sum_{i=1}^{n} X_i^k\right)^2\right) \\ &= \nu_2 - \frac{1}{n^2} E\left(\sum_{i=1}^{n} X_i^2 + 2\sum_{i<j} X_i X_j\right) \\ &= \nu_2 - \frac{1}{n^2}\left[\sum_{i=1}^{n} E\left(X_i^2\right) + 2\sum_{i<j} E(X_i)E(X_j)\right] \\ &= \nu_2 - \frac{1}{n^2}\left[n\nu_2 + 2\,\frac{n(n-1)}{2}\,\nu_1^2\right] = \nu_2 - \frac{1}{n}\nu_2 - \frac{n-1}{n}\nu_1^2 \\ &= \frac{n-1}{n}\left(\nu_2 - \nu_1^2\right) = \frac{n-1}{n}\,\sigma^2. \end{aligned}$$

□

Remark 4.12.
1. For large samples, i.e. when $n \to \infty$, $\overline{X}$ and $\overline{\mu}_2$ are uncorrelated.
2. If X has a symmetric distribution, then $\mu_3 = 0$ and, hence, $\overline{X}$ and $\overline{\mu}_2$ are uncorrelated for every $n \in \mathbb{N}$.
3. As before, one can show that under the assumptions of Proposition 4.11,

the sequence

$$Z_n = \frac{\overline{\mu}_2 - \sigma^2}{\sqrt{\frac{\mu_4 - \sigma^4}{n}}}$$

converges in distribution to a standard normal variable, as $n \to \infty$.
4. Notice that the sample central moment of order 2 is the first statistic whose expected value is not the corresponding population function, in this case the theoretical variance. This is the motivation for the next definition.

Definition 4.13. *The statistic*

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2 \tag{33}$$

*is called the **sample variance*** (var) *and its value is* $\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2$.

The statistic $s = \sqrt{s^2}$ *is called the **sample standard deviation*** (std).

Remark 4.14. Notice that the sample central moment of order 2 is no longer equal to the sample variance, as we are used to. In fact, we have

$$s^2 = \frac{n}{n-1}\, \overline{\mu}_2.$$

Then, by Proposition 4.11, we have for the sample variance

$$\begin{aligned} E\left(s^2\right) &= \mu_2 = \sigma^2, \\ V\left(s^2\right) &= \frac{1}{n(n-1)}[(n-1)\mu_4 - (n-3)\sigma^4], \\ \operatorname{cov}(\overline{X}, s^2) &= \frac{1}{n}\, \mu_3. \end{aligned} \tag{34}$$

Sample Distribution Function

Thus far, we have been able to define sample functions that mimicked their theoretical correspondents (mean, moments, variance) and, hopefully, will provide good inferential estimates for the entire population. The ultimate goal of Statistics is to derive the probability distribution that generated a sample from the sample itself, i.e. to define a sample function that gives

some idea of the cumulative distribution function of a characteristic, relative to the entire population. The idea is suggested by the shape of the cumulative distribution function of discrete random variables.

Definition 4.15. *Let X be a characteristic and $X_1, \ldots, X_n$ sample variables for a random sample of size n. The* ***sample distribution function*** *or* ***empirical distribution function*** (cdfplot) *is the sample function $\overline{F}_n : \mathbb{R} \to \mathbb{R}$ defined by*

$$\overline{F}_n(x) = \frac{1}{n} I(X_i \leq x) = \frac{\text{card}\{X_i \mid X_i \leq x\}}{n}, \tag{35}$$

where $I(A)$ is the indicator of event A, and its value is $\dfrac{\text{card}\{x_i \mid x_i \leq x\}}{n}$.

Remark 4.16.
1. So, the sample distribution function at a value x is given by

$$\overline{F}_n(x) = \frac{\text{number of sample elements } x_i \leq x}{n}.$$

Defining it this way makes it an excellent tool for measuring how faithful the sample is to the probability distribution: where observations are densely packed, this function grows rapidly, which is exactly what is expected from the true distribution function (for where the distribution function grows rapidly, the probability density−its derivative, is large, which is propitious to a high concentration of observations), while observations that are few and far between happen in regions of low probability density.
2. Assuming the sample data $x_1, \ldots, x_n$ are sorted in increasing order, a more explicit computational formula for the sample distribution function is

$$\overline{F}_n(x) = \begin{cases} 0, & \text{if } x < x_1 \\ \dfrac{i}{n}, & \text{if } x_i \leq x < x_{i+1},\ i = \overline{1, n-1} \\ 1, & \text{if } x \geq x_n. \end{cases}$$

Thus $\overline{F}_n$ presents similar properties to those of a cumulative distribution function of a discrete random variable:

– it is a step function;
– it monotonically increases from 0 to 1;
– it is constant on semi-open intervals $[x_i, x_{i+1})$;
– its limits at $\pm\infty$ are 1 and 0, respectively.
In addition, here the height of each "step" is $\dfrac{1}{n}$.

3. The sample distribution function can also be viewed as a random variable (since it *is* a sample function). If F denotes the cumulative distribution function of the characteristic X, then for each $x \in \mathbb{R}$, $\overline{F}_n(x)$ is a discrete random variable with probability distribution function

$$\overline{F}_n(x) \begin{pmatrix} \dfrac{i}{n} \\ \\ C_n^i \left(F(x)\right)^i \left(1 - F(x)\right)^{n-i} \end{pmatrix}_{i=\overline{0,n}}.$$

Now that we have seen the similarities between a cumulative distribution function and a sample distribution function, the question that naturally arises is how much does the latter resemble the former, how well and in what sense, does it approximate it. Of course, since these are random variables, convergence of such types of variables should be considered. The Weak Law of Large Numbers can be used to show that

$$\overline{F}_n(x) \xrightarrow{p} F(x),$$

for every fixed $x \in \mathbb{R}$. But an even stronger convergence result holds:

Theorem 4.17 (Glivenko-Cantelli)**.** *Let X be a characteristic with cumulative distribution function F and $X_1, \ldots, X_n$ sample variables for a random sample of size n, with sample distribution function $\overline{F}_n$. Let*

$$D_n = \sup_{x \in \mathbb{R}} |\overline{F}_n(x) - F(x)|. \tag{36}$$

Then

$$P\left(\lim_{n\to\infty} D_n = 0\right) = 1,$$

i.e. the sample distribution function converges almost surely to the cumulative distribution function.

Kolmogorov strengthened this result, by effectively providing the rate of this convergence. A random variable is said to follow the **Kolmogorov distribution**, if its cumulative distribution function is given by

$$K(x) = 1 - 2\sum_{k=1}^{\infty}(-1)^{k-1}e^{-2k^2x^2} = \frac{\sqrt{2\pi}}{x}\sum_{k=1}^{\infty}e^{-\frac{(2k-1)^2\pi^2}{8x^2}}, \qquad (37)$$

for all $x > 0$ and 0, otherwise. The function (37) is known as *Kolmogorov's function* and its values can be found in tables.

Theorem 4.18 (Kolmogorov)**.** *Assume the hypotheses of Theorem 4.17 are satisfied and further, assume that F is continuous. Then*

$$\lim_{n\to\infty} P\left(\sqrt{n}D_n \leq x\right) = K(x),$$

for all $x > 0$, i.e. the variable $\sqrt{n}D_n$ converges in distribution to the Kolmogorov distribution.

4.2 Properties of Sample Functions

Sample functions are very important in inferential Statistics, since they represent the only "real" information we have about a population. The goal is to be able to make "predictions" on a population characteristic, based on this information that a sample provides and also to measure (in terms of probability) how good those predictions are. Results such as Proposition 4.5 and Corollary 4.6 can be very useful in that sense. We present next more such results (without proof), that will be used in the next two chapters in making inferences about population characteristics.

Let X be a characteristic of a population from which a random sample of size n is drawn and let $X_1, \ldots, X_n$ be the sample variables.

Proposition 4.19. *Assume $X \in N(0,1)$ and let*

$$U_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}X_i = \sqrt{n}\,\overline{X} \quad \textit{and} \quad V_n = \sum_{i=1}^{n}(X_i - \overline{X})^2 = (n-1)\,s^2.$$

Then $U_n \in N(0,1)$ and $V_n \in \chi^2(n-1)$.

Proposition 4.20. *Assume $X \in N(\mu, \sigma)$ and let*

$$U_n = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \quad \textit{and} \quad V_n = \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \overline{X})^2 = \frac{(n-1)\, s^2}{\sigma^2}.$$

Then $U_n \in N(0,1)$ and $V_n \in \chi^2(n-1)$.

Proposition 4.21. *Assume $X \in N(\mu, \sigma)$ and let*

$$T = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}}.$$

Then $T \in T(n-1)$.

It will be necessary sometimes to compare characteristics of two populations. For that, we will need results on sample functions referring to both collections. Assume we have two characteristics $X_{(1)}$ and $X_{(2)}$, relative to two populations. We draw from both populations random samples of sizes n_1 and n_2, respectively. Denote the two sets of random variables by

$$X_{11}, \ldots, X_{1n_1} \text{ and } X_{21}, \ldots, X_{2n_2}.$$

Denote the sample means and sample variances by

$$\overline{X}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} X_{1i}, \quad \overline{X}_2 = \frac{1}{n_2} \sum_{j=1}^{n_2} X_{2j}$$

and

$$s_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} \left(X_{1i} - \overline{X}_1\right)^2, \quad s_2^2 = \frac{1}{n_2 - 1} \sum_{j=1}^{n_2} \left(X_{2j} - \overline{X}_2\right)^2.$$

In addition, denote by

$$s_p^2 = \frac{\sum_{i=1}^{n_1} \left(X_{1i} - \overline{X}_1\right)^2 + \sum_{j=1}^{n_2} \left(X_{2j} - \overline{X}_2\right)^2}{n_1 + n_2 - 2} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

the **pooled variance** of the two samples, i.e. a variance that considers the sample data from both samples.

Proposition 4.22. *Assume* $X_{(1)} \in N(\mu_1, \sigma_1)$ *and* $X_{(2)} \in N(\mu_2, \sigma_2)$ *are independent and let*

$$Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \quad \textit{and} \quad T = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}.$$

Then $Z \in N(0,1)$ *and* $T \in T(n)$*, where*

$$\frac{1}{n} = \frac{c^2}{n_1 - 1} + \frac{(1-c)^2}{n_2 - 1} \quad \textit{and} \quad c = \frac{\dfrac{s_1^2}{n_1}}{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}.$$

Proposition 4.23. *Assume* $X_{(1)} \in N(\mu_1, \sigma)$ *and* $X_{(2)} \in N(\mu_2, \sigma)$ *are independent and let*

$$T = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}.$$

Then $T \in T(n_1 + n_2 - 2)$.

Proposition 4.24. *Assume* $X_{(1)} \in N(\mu_1, \sigma_1)$ *and* $X_{(2)} \in N(\mu_2, \sigma_2)$ *are independent and let*

$$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}.$$

Then $F \in F(n_1 - 1, n_2 - 1)$.