PART II. STATISTICS

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 (a set of observed units collected from 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 (estimates) are made about the entire population, as well as measurements of the degree of accuracy of the estimates.

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**.

Sometimes discrepancies occur between a sample and its underlying population.
Sampling errors are caused simply by the fact that only a portion of the entire population is observed. For most statistical procedures, sampling errors decrease (and converge to zero) if the sample size is appropriately increased.
Non-sampling errors are produced by inappropriate sampling designs or wrong statistical techniques. No statistical procedures can save a poorly collected sample!

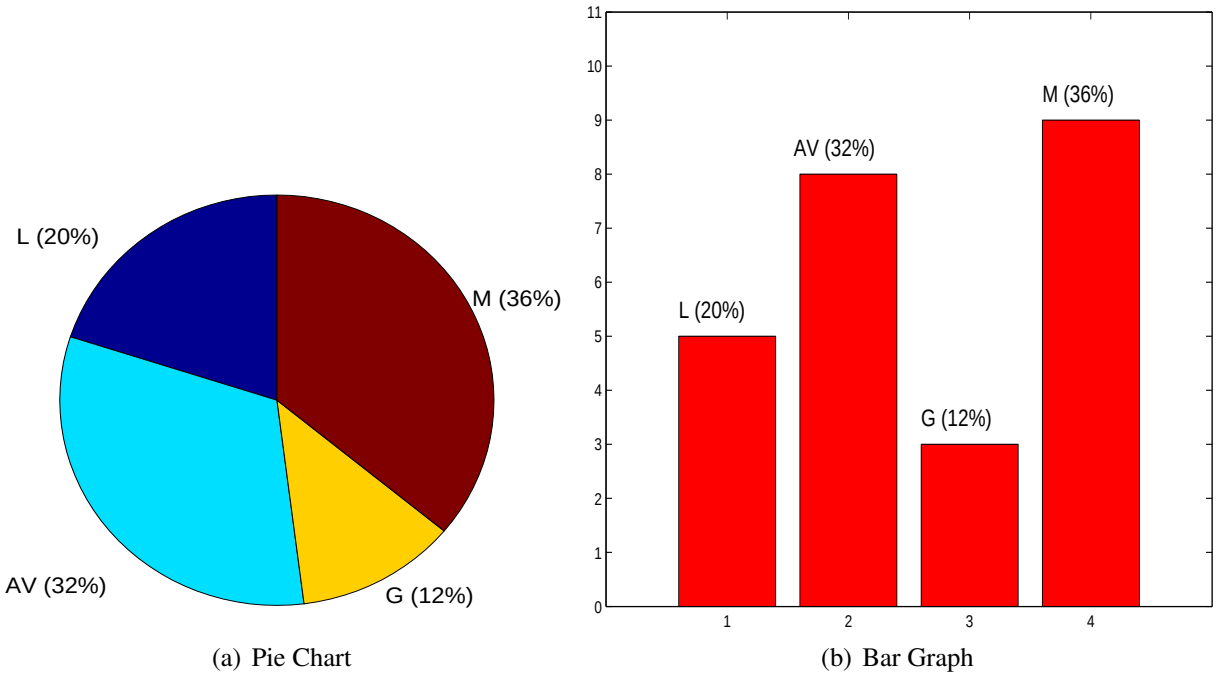
1.3 Graphical Display of Data, Frequency Distribution Tables, Histograms

"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 non-repetitive, 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** $x_i = \frac{c_{i-1} + c_i}{2}$, 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*

$$n = 1 + \frac{10}{3}\log_{10} N, \tag{1.1}$$

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

$$l = \frac{8}{100}\left(x_{\max} - x_{\min}\right). \tag{1.2}$$

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 (pdf) 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 a group 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 (1.2), 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.

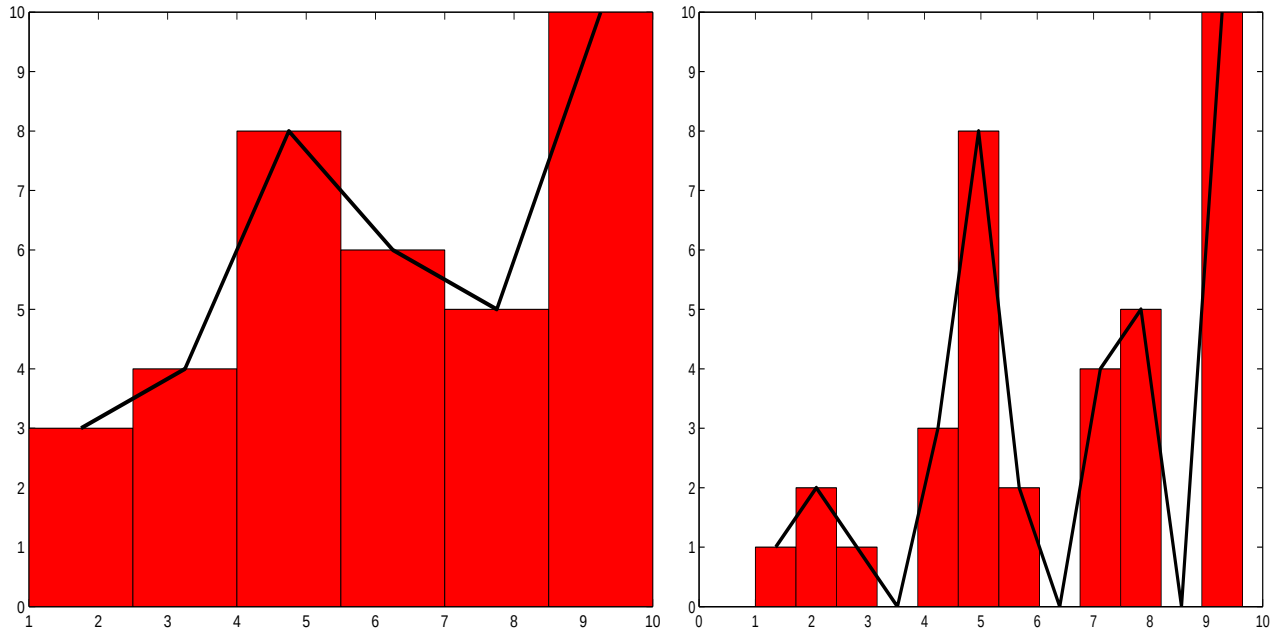
No	Class	Mark	Freq.	C. Freq.	R. Freq.	R. C. Freq.
1	[1.00 , 2.50)	1.75	3	3	8%	8%
2	[2.50 , 4.00)	3.25	4	7	11%	19%
3	[4.00 , 5.50)	4.75	8	15	22%	41%
4	[5.50 , 7.00)	6.25	6	21	17%	58%
5	[7.00 , 8.50)	7.75	5	26	14%	72%
6	[8.50 , 10.00]	9.25	10	36	28%	100%

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

No	Class	Mark	Freq.	C. Freq.	R. Freq.	R. C. Freq.
1	[1.00 , 1.72)	1.36	1	1	3%	3%
2	[1.72 , 2.44)	2.08	2	3	6%	9%
3	[2.44 , 3.16)	2.80	1	4	3%	12%
4	[3.16 , 3.88)	3.52	0	4	0%	12%
5	[3.88 , 4.60)	4.24	3	7	8%	20%
6	[4.60 , 5.32)	4.96	8	15	22%	42%
7	[5.32 , 6.04)	5.68	2	17	6%	48%
8	[6.04 , 6.76)	6.40	0	17	0%	48%
9	[6.76 , 7.48)	7.12	4	21	11%	59%
10	[7.48 , 8.20)	7.84	5	26	14%	73%
11	[8.20 , 8.92)	8.56	0	26	0%	73%
12	[8.92 , 10]	9.46	10	36	27%	100%

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



(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*

$$\overline{x}_a = \frac{1}{N} \sum_{i=1}^{N} x_i. \tag{2.1}$$

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

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

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*

$$\overline{x}_g = \sqrt[N]{x_1 \ldots x_N}. \tag{2.2}$$

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

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

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 absolute deviations, emphasized by the arithmetic mean.

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

$$\overline{x}_h = \frac{N}{\sum_{i=1}^{N} \frac{1}{x_i}}. \tag{2.3}$$

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 $x_1, \dots, x_N$, the well-known *means inequality* holds:

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

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

One disadvantage of the sample mean is its *sensitivity to extreme observations*. The next measure of location is the *median*, which is much less affected by extreme values than the mean.

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 that it is exceeded by at most a half of observations and is preceded by at most a half of observations.*

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. Notice from the wording of the definition that the mode may not be unique. A distribution can have one mode − **unimodal**, two modes − **bimodal**, three modes − **trimodal**, or more − **multimodal**.
When the pdf of a continuous distribution has multiple local maxima, it is common to refer to *all* of the local maxima as modes of the distribution.
If every value occurs only once in a sample, we say that there is **no mode**.

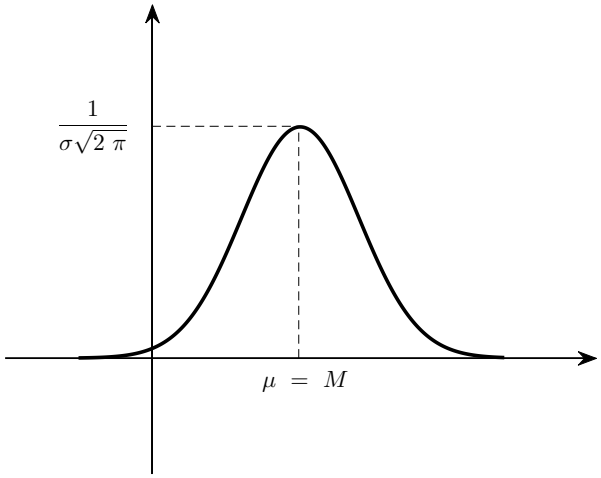
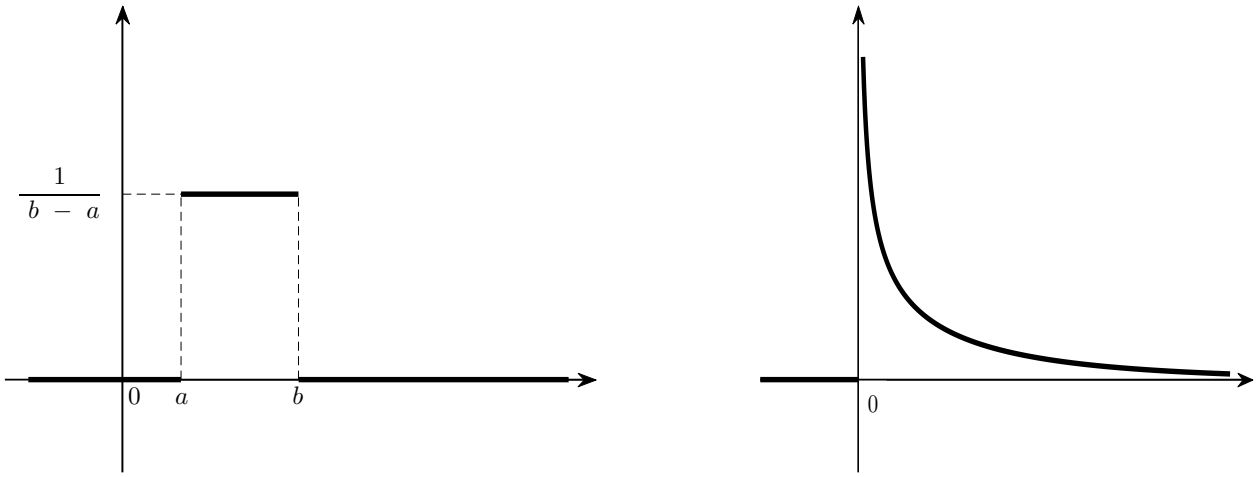


Fig. 3: Normal Distribution Mode

For data drawn from symmetric distributions, we have

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

This is true, for instance, for the Normal distribution which is unimodal (see Figure 3). For a Uniform $U(a, b)$ distribution, *all* values in the interval $[a, b]$ are modes, while the $\chi^2(1)$ distribution (with $\nu = 1$ degree of freedom) has no mode (see Figure 4).



(a) Uniform Distribution (multimodal)

(b) χ^2 Distribution (no mode)

Fig. 4: Modes

In general,

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

This empirical formula was given by K. Pearson.

2.2 Measures of Variability

Once we have located the central values of a set of data, it is important to measure the *variability*, whether the data values are tightly clustered or spread out. At the heart of Statistics lies variability: measuring it, reducing it, distinguishing random from real variability, identifying the various sources of real variability and making decisions in the presence of it. We need to know how "unstable" the data is and how much the values differ from its average or from other middle values. These numbers will have small values for closely grouped data (little variation) and larger values for more widely spread out data (large variation).

The measures of variation will also help us assess the reliability of our estimates and the accuracy of our forecasts.

Consider the primary data $X = \{x_1, \ldots, x_N\}$. The first three 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 mean of the absolute value of the deviations from the mean, i.e. the value*

$$MAD_1 = \frac{1}{N}\sum_{i=1}^{N} |x_i - \overline{x}|.$$

*The **median absolute deviation*** (mad) *of X is the median of the absolute value of the deviations from the median, i.e. the value*

$$MAD_2 = \text{median}\{|x_i - x_{me}|\}.$$

Like the median, the median absolute deviation is not influenced by extreme values, whereas the mean absolute deviation is.

Next, following the idea behind the definition of the median, we define values that divide the data into certain percentages. We simply replace 0.5 in its definition by some probability $0 < p < 1$.

Definition 2.12. *Let X be a set of data sorted increasingly, $p \in (0, 1)$ and $k = 1, 2, \ldots, 99$.*

(1) *A **sample p-quantile*** (prctile) *is any number that exceeds at most $100p\%$ of the sample and is exceeded by at most $100(1 - p)\%$ of the sample.*

(2) *A **k-percentile** P_k is a $(k/100)$-quantile. So, P_k exceeds at most $k\%$ and is exceeded by at most $(100 - k)\%$ of the data*

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

$$Q_1 = P_{25}, \;\; Q_2 = P_{50} = \overline{M} \;\text{ and }\; Q_3 = P_{75}.$$

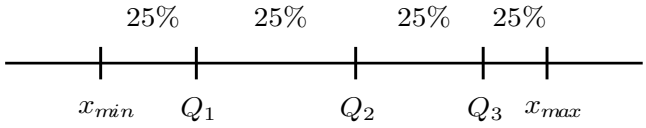


Fig. 5: Quartiles

Definition 2.13. *Let X be a set of sorted data with quartiles Q_1, Q_2 and Q_3.*

(1) *The **interquartile range*** (iqr) *is the difference between the third and the first quartile*

$$IQR = Q_3 - Q_1. \tag{2.4}$$

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

$$IQD = \frac{IQR}{2} = \frac{Q_3 - Q_1}{2}. \tag{2.5}$$

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

$$IQDC = \frac{IQD}{x_{me}} = \frac{Q_3 - Q_1}{2Q_2}. \tag{2.6}$$

Remark 2.14.
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.

Outliers

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 but one are between 0 and 1, a value of 1000 would surely seem out of place!

```
>> x=[rand(10,1); 1000]
x =

   1.0e+03 *

    0.0007
    0.0000
    0.0003
    0.0000
    0.0001
    0.0008
    0.0007
    0.0003
    0.0010
    0.0000
    1.0000
```

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.

```
>> mean(x)

ans =

   91.2707

>> median(x)
```

```
ans =

    0.3171
```

Once the presence of such outliers is detected, it is suggested that sample statistics be computed both *with* and *without* the outliers.

```
>> x = x(1:end-1)

x =

    0.7000
         0
    0.3000
         0
    0.1000
    0.8000
    0.7000
    0.3000
    1.0000
         0

>> mean(x)

ans =

    0.3900

>> median(x)

ans =

    0.3000
```

Thus the problem of detecting and locating an outlier is an important part of any statistical data analysis process.

How to classify a value as being "extreme"? 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" (the "3σ rule" is an application of Chebyshev's inequality and states that most of the values that any random variable takes, at least 89%, lie within 3 standard deviations away from the mean) and would work well for unimodal and symmetrical distributions.

A more general approach, that works for skewed data, is to consider an outlier 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 2.15. Consider the following set of data

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

We sort them in increasing order:

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

We have:

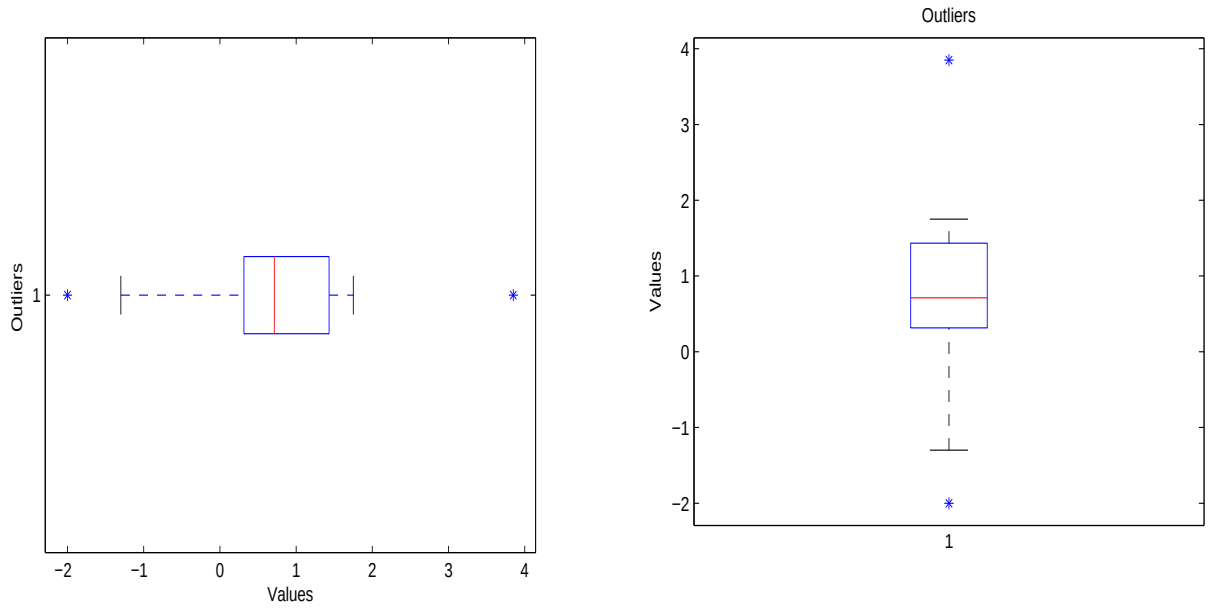
$$\begin{aligned} Q1 &= 0.3150, \\ Q2 &= 0.7133, \\ Q3 &= 1.4323, \\ IQR &= 1.1173, \end{aligned}$$

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

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

so observations outside the interval $[-1.3610, 3.1082]$ are considered outliers. In this case, there are

two such values, -2 and 3.85. The data (boxplot) is displayed graphically in Figure 6.



(a) horizontally (b) vertically

Fig. 6: Quartiles, Interquartile Range, Outliers

Boxplots are also very useful when we want to compare data from different samples (see Figure 7). We can compare the interquartile ranges, to examine how the data is dispersed between each sample. The longer the box, the more dispersed the data.

Definition 2.16.

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

$$\overline{\nu}_k = \frac{1}{N}\sum_{i=1}^{N} x_i^k, \quad \overline{\nu}_k = \frac{1}{N}\sum_{i=1}^{n} f_i x_i^k, \tag{2.7}$$

for primary and for grouped data, respectively.

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

$$\overline{\mu}_k = \frac{1}{N}\sum_{i=1}^{N} (x_i - \overline{x})^k, \quad \overline{\mu}_k = \frac{1}{N}\sum_{i=1}^{n} f_i (x_i - \overline{x})^k \tag{2.8}$$

for primary and for grouped data, respectively.

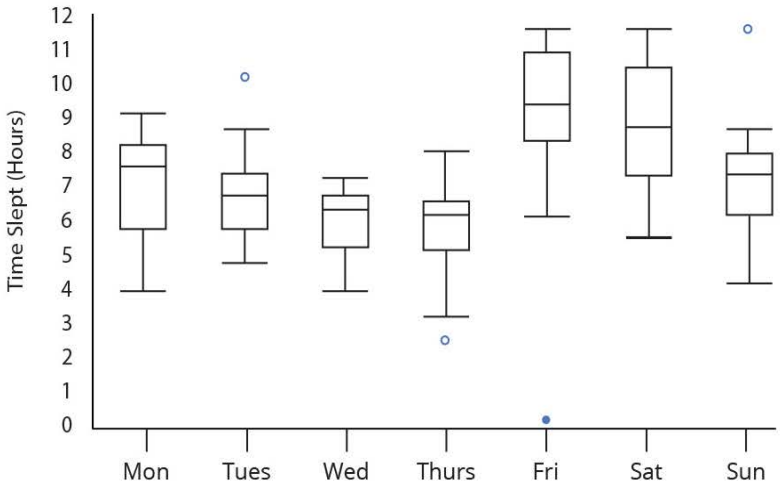


Fig. 7: Multiple boxplots

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

$$\overline{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2, \;\; \overline{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{n} f_i(x_i - \overline{x})^2 \tag{2.9}$$

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

Remark 2.17.

1. A more efficient computational formula for the variance is

$$\overline{\sigma}^2 = \frac{1}{N}\left(\sum_{i=1}^{N} x_i^2 - N\Big(\sum_{i=1}^{N} x_i\Big)^2\right) = \frac{1}{N}\left(\sum_{i=1}^{N} x_i^2 - N\overline{x}^2\right), \tag{2.10}$$

which follows straight from the definition.

2. We will see later that when the data represents a sample (not the entire population), a better formula for the variance is

$$\begin{aligned} s^2 &= \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2 &&= \frac{1}{N-1}\Big(\sum_{i=1}^{N} x_i^2 - N\overline{x}^2\Big), \\ s^2 &= \frac{1}{N-1}\sum_{i=1}^{n} f_i(x_i - \overline{x})^2 &&= \frac{1}{N-1}\Big(\sum_{i=1}^{N} f_i x_i^2 - N\overline{x}^2\Big), \end{aligned} \tag{2.11}$$

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 (2.9) to compute the variance of a set of data that represents a population and (2.11) for the variance of a

sample.

Definition 2.18. *The **coefficient of variation** is the value*

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

Remark 2.19.
1. The coefficient of variation is also known as the **relative standard deviation (RSD)**.
2. It 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 used in fields such as analytical chemistry, engineering or physics when doing quality assurance studies. It is also widely used in 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 trade off.

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, called the *response* variable and one or more independent variables, called *predictor* variables (or, simply, predictors). 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 in Table 5.

$X \setminus Y$	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*

$$\overline{\nu}_{k_1k_2} = \frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}x_i^{k_1}y_j^{k_2}. \tag{3.1}$$

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

$$\overline{\mu}_{k_1k_2} = \frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}(x_i-\overline{x})^{k_1}(y_j-\overline{y})^{k_2}, \tag{3.2}$$

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:

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

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

(1) *The **covariance*** ($\boxed{\text{cov}}$) *of (X, Y) is the value*

$$\text{cov}(X, Y) = \overline{\mu}_{11} = \frac{1}{N} \sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij}(x_i - \overline{x})(y_j - \overline{y}). \tag{3.3}$$

(2) *The **correlation coefficient*** ($\boxed{\text{corrcoef}}$) *of (X, Y) is the value*

$$\overline{\rho} = \overline{\rho}_{XY} = \frac{\text{cov}(X, Y)}{\sqrt{\overline{\mu}_{20}}\sqrt{\overline{\mu}_{02}}} = \frac{\overline{\mu}_{11}}{\overline{\sigma}_X \overline{\sigma}_Y}. \tag{3.4}$$

These two notions have been mentioned before, 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 $\text{cov}(X, Y) = \overline{\nu}_{11} - \overline{x} \cdot \overline{y}$.

The correlation coefficient is then

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

and, as before, it satisfies the inequality

$$-1 \leq \overline{\rho} \leq 1 \tag{3.5}$$

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. In other words, we should consider *conditional means*, defined similarly to regular means, only taking into account the condition.

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*

$$\overline{y}_i = \overline{y}(x_i) = \frac{1}{f_{i\cdot}} \sum_{j=1}^{n} f_{ij} y_j,\ i = \overline{1, m}. \tag{3.6}$$

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

$$\overline{x}_j = \overline{x}(y_j) = \frac{1}{f_{\cdot j}} \sum_{i=1}^{m} f_{ij} x_i,\ j = \overline{1, n}. \tag{3.7}$$

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 $x = g(y)$ 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) (the sum of the squares of the differences between the responses y_j and their fitted values $y(x_i)$, each counted

with the corresponding frequency)

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

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.

$$-2\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}\Big(y_j - f(x_i; a_1, \ldots, a_s)\Big)\frac{\partial f(x_i; a_1, \ldots, a_s)}{\partial a_k} = 0, \tag{3.8}$$

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

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

is minimum. The system (3.8) becomes

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

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

$$y - \overline{y} = \overline{\rho}\,\frac{\overline{\sigma}_Y}{\overline{\sigma}_X}\,(x - \overline{x}) \tag{3.9}$$

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

$$x - \overline{x} = \overline{\rho}\,\frac{\overline{\sigma}_X}{\overline{\sigma}_Y}\,(y - \overline{y})\,. \tag{3.10}$$

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).

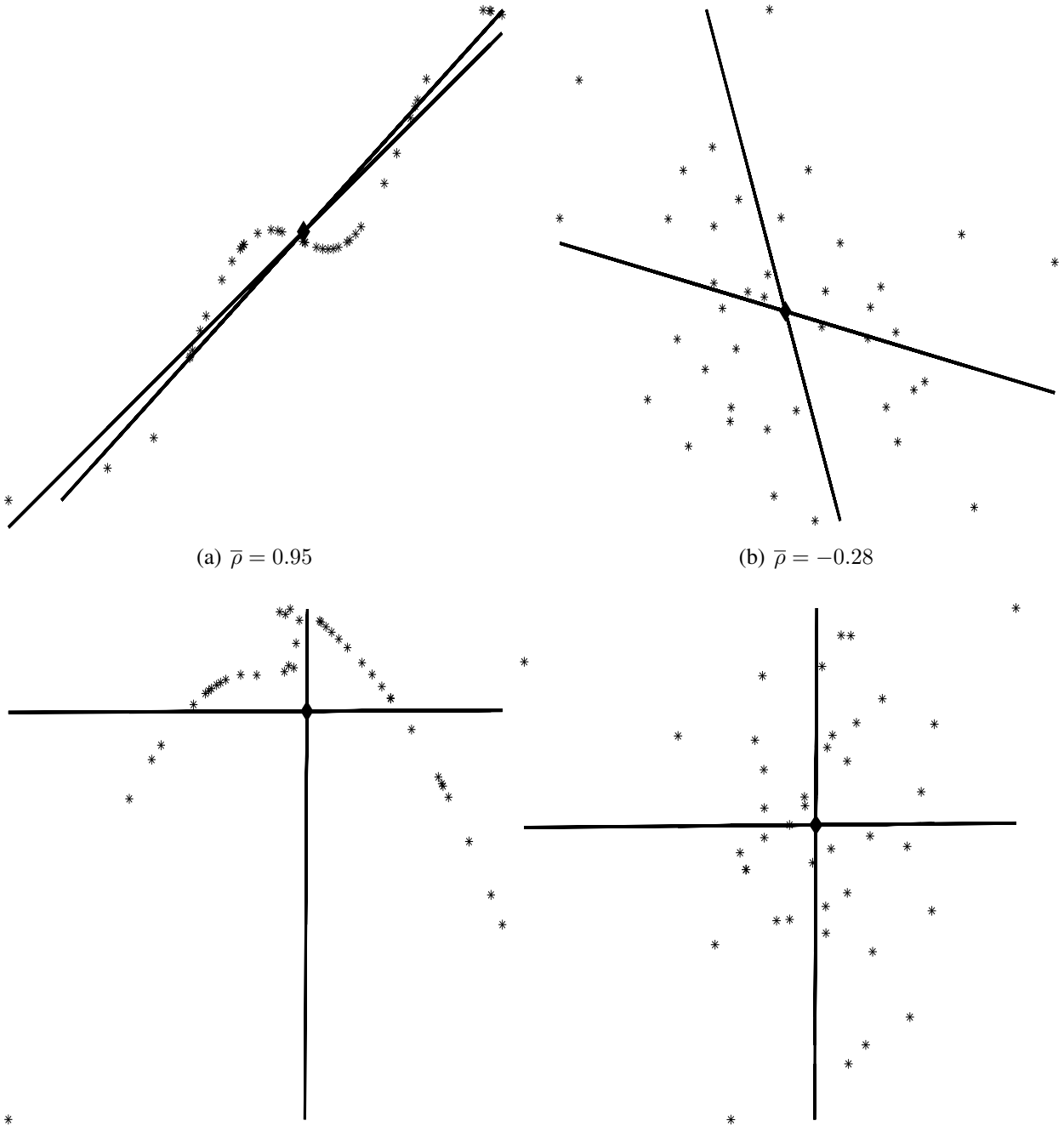
2. The slope $\overline{a}_{Y|X} = \overline{\rho}\,\dfrac{\overline{\sigma}_Y}{\overline{\sigma}_X}$ of the line of regression of Y on X is called the *coefficient of regression* of Y on X. Similarly, $\overline{a}_{X|Y} = \overline{\rho}\,\dfrac{\overline{\sigma}_X}{\overline{\sigma}_Y}$ is the coefficient of regression of X on Y and

$$\overline{\rho}^2 = \overline{a}_{Y|X}\;\overline{a}_{X|Y}.$$

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



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

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

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

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

Fig. 8: Scattergram, Lines of Regression and Centroid

Example 3.8. Let us examine the situations graphed in Figure 8.

- In Figure 8(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 8(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 8(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 $Y = -X^2 + \sin\left(\frac{1}{X}\right)$. Notice also, that the two lines of regression are perpendicular.

- Finally, in Figure 8(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.

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 $y = \dfrac{a}{x} + b$.