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.

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 (predictions, approximations).

Historically, descriptive Statistics was developed first, dealing with the "raw" data that people had to handle every day. As that task became increasingly difficult, a more scientific approach of Statistics was needed. Modern Statistics, as a rigorous scientific discipline, traces its roots back to the late 1800's and F. Galton and K. Pearson.

A new trend in modern Statistics is **Exploratory Data Analysis (EDA)**. This new area of Statistics was promoted by John Tukey beginning in the 1970's. He encouraged statisticians to *explore* the data, often using statistical graphics and other data visualization methods, and possibly formulate hypotheses that could lead to new data collection and experiments. With the ready availability of computing power and expressive data analysis software, EDA has evolved constantly in recent decades, by means of the rapid development of new technology and access to more and bigger data.

1 Basic Concepts. Terminology

- A **population** is a set of individuals, objects, items or measurements of interest, 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 **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* (categorical, a certain trait). From the probabilistic point of view, a numerical characteristic is a random variable.

- A numerical characteristic is called a **parameter**, if it refers to an entire population and a **statistic** or **sample function**, if it refers just to a sample. Populations are characterized by *parameters* - usually unknown, which are to be estimated based on *statistics* - known from the sample(s) collected.
- 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.
- Depending on the goal of a data analysis project, the data gathered can be of several types:
  - **discrete**, data that can take on only a discrete set of values (data that can be counted);
  - **continuous**, data that can take on any value in an (possibly infinite) interval (data that can be measured);
  - **categorical**, data that can take on only a specific set of values representing a set of possible categories;
  - **binary**, a special case of categorical data with just two categories of values (0/1, yes/no, true/false);
  - **ordinal**, categorical data that has an explicit ordering.

2 Data Collection

2.1 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. A random sample must also be representative for the population it was drawn from (the structure of the sample must be similar to the structure of the population).

Other types of samples may be considered: **systematic** sample, **stratified** sample, **quota** sample, **cluster** sample, etc.

Throughout the remaining chapters, we will only consider **simple random sampling**, i.e. a sampling design where units are collected from the entire population independently of each other, all being equally likely to be sampled. Observations collected by means of a simple random sampling design are **iid (independent, identically distributed)** random variables.

2.2 Sampling and Non-Sampling Errors

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!

3 Graphical Display of Data

"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 3.1. 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). Pie charts are shown in Figure 1 and bar graphs in Figure 2.

3.1 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.

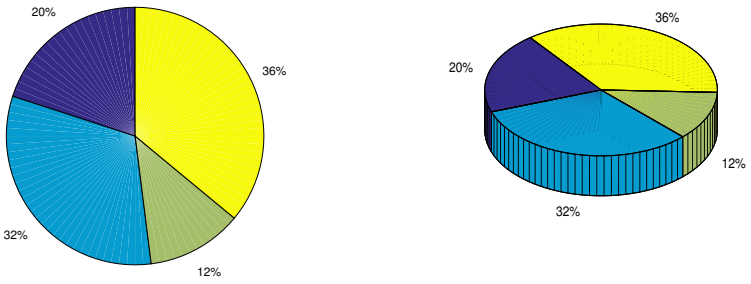


Fig. 1: Pie Charts

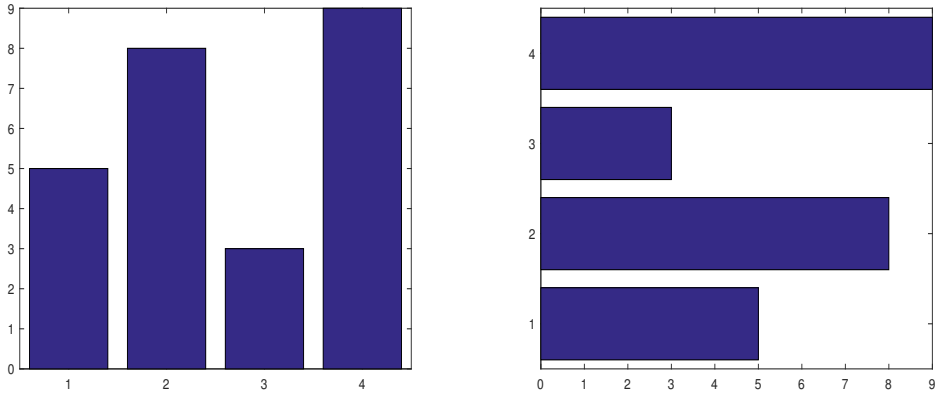


Fig. 2: Bar graphs

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 3.2.

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).

Determining the number of classes

There isn't an "optimal" way of choosing the number of classes (bins) to group data. But in general,

- there should not be too few or too many classes;
- their number may increase with the sample size;
- they should be chosen to make the frequency distribution table (and then, further, its visual counterparts, the histogram, the frequency polygon, the stem-and-leaf plot) informative, so that we can notice patterns, shapes, outliers, etc.

We can start with $n = 10$ classes (most software have that as the implicit number), see what information we get and then decide whether to increase or decrease the number of bins.

There is, also, a customary procedure (empirical formula) of determining the number of classes, known as *Sturges' rule*

$$n = 1 + \frac{10}{3}\log_{10} N, \tag{3.1}$$

where N is the sample size. Then it follows that

$$l = \frac{x_{\max} - x_{\min}}{n}.$$

Once we determined n and l, we have

$$c_i = x_{\min} + i \cdot l,\ i = \overline{0, n}.$$

Example 3.3. To evaluate effectiveness of a processor for a certain type of tasks, the random variable X, the CPU time of a job, is studied. The following data represent the CPU times for $n = 30$ randomly chosen jobs (in seconds):

70	36	43	69	82	48	34	62	35	15
59	139	46	37	42	30	55	56	36	82
38	89	54	25	35	24	22	9	56	19

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

$$
\begin{array}{cccccccccc}
9 & 15 & 19 & 22 & 24 & 25 & 30 & 34 & 35 & 35 \\
36 & 36 & 37 & 38 & 42 & 43 & 46 & 48 & 54 & 55 \\
56 & 56 & 59 & 62 & 69 & 70 & 82 & 82 & 89 & 139
\end{array}
$$

There are $N = 30$ observations, with $x_{\min} = 9$ and $x_{\max} = 139$.

Since there are very few repetitions, an ungrouped frequency distribution table would not tell us much.

Let us group the data into classes of the same length. With $n = 10$ bins, we have a class width of $l = 13$, whereas with Sturges' rule, we get $n = 5.9237 \approx 6, \;\; l \approx 21.7$.

The grouped frequency tables are shown in Tables 2 and 3. We have also included the relative and cumulative frequencies.

No	Class	Mark	Freq.	C. Freq.	R. Freq.	R. C. Freq.
1	[9, 22]	15.5	4	4	13%	13%
2	(22, 35]	28.5	6	10	20%	33%
3	(35, 48]	41.5	8	18	27%	60%
4	(48, 61]	54.5	5	23	17%	77%
5	(61, 74]	67.5	3	26	10%	87%
6	(74, 87]	80.5	2	28	7%	94%
7	(87, 100]	93.5	1	29	3%	97%
8	(100, 113]	106.5	0	29	0%	97%
9	(113, 126]	119.5	0	29	0%	97%
10	(126, 139]	132.5	1	30	3%	100%

Table 2: Example 3.3, Grouped frequency distribution table with $n = 10$ classes

No	Class	Mark	Freq.	C. Freq.	R. Freq.	R. C. Freq.
1	[9, 30.7)	19.85	7	7	23%	23%
2	[30.7, 52.4)	41.55	11	18	37%	60%
3	[52.4, 74.1)	63.25	8	26	27%	87%
4	[74.1, 95.8)	84.95	3	29	10%	97%
5	[95.8, 117.5)	106.65	0	29	0%	97%
6	[117.5, 139)	128.35	1	30	3%	100%

Table 3: Example 3.3, Grouped frequency distribution table with $n = 6$ classes

Remark 3.4. 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.

3.2 Histograms and Frequency Polygons

When data is grouped into classes, the best way to visualize the frequency distribution is by constructing a **histogram** (in Matlab hist/histogram). 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.

A histogram shows the shape of a pdf (probability distribution/density function) or pmf (probability mass function) of data, checks for homogeneity, and suggests possible outliers.

A **frequency histogram** consists of columns, one for each class (bin), whose height is determined by the number of observations in the bin (i.e, the class frequency).

A **relative frequency histogram** has the same shape but a different vertical scale. Its column heights represent the proportion of all data that appeared in each bin.

If relative frequencies are considered, 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).

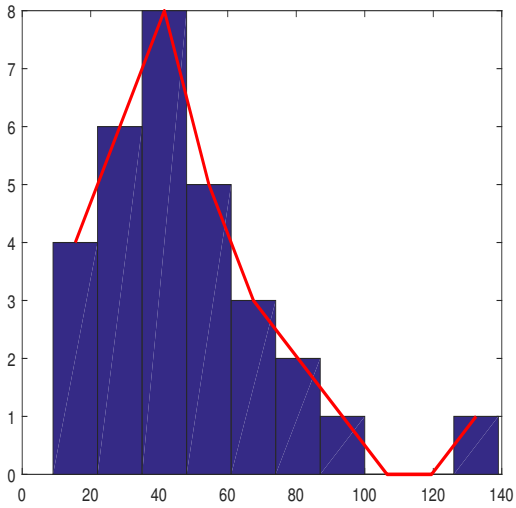
Example 3.5. Let us consider again the data in Example 3.3, the CPU times (in seconds) for $N = 30$ randomly chosen jobs:

70	36	43	69	82	48	34	62	35	15
59	139	46	37	42	30	55	56	36	82
38	89	54	25	35	24	22	9	56	19

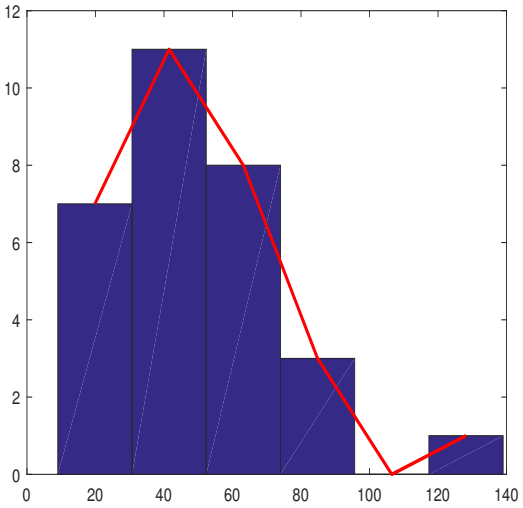
We constructed the grouped frequency distribution tables for these data for $n = 10$ and for $n = 6$ classes. Figure 3 shows the corresponding histogram and frequency polygon for grouped data ((a) and (b)). Also in Figure 3, we show histograms for $n = 4$ and $n = 12$ bins, respectively. It is obvious that $n = 4$ is too small and $n = 12$ is too large for the number of bins. The values $n = 6$ and $n = 10$ seem to be the best (in terms of the information they provide), especially $n = 10$.

For 10 classes, let us take a closer look (see Figure 4). What information can we draw from these histograms?

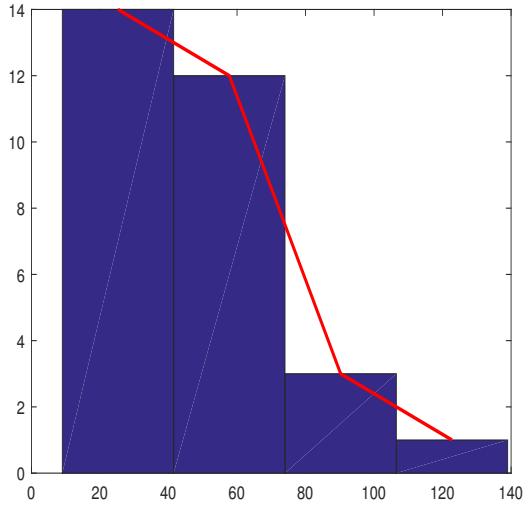
- the continuous distribution (continuous, because time varies *continuously*) of the CPU times is not symmetric, it is skewed to the right, as we see 5 columns to the right of the highest column and only 2 columns to the left;



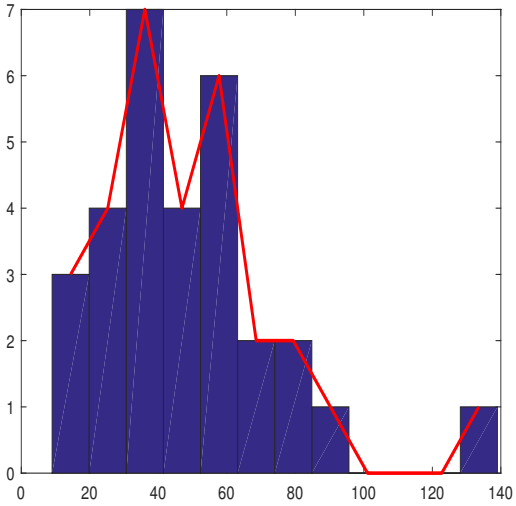
(a) $n = 10$ bins



(b) $n = 6$ bins



(c) $n = 4$ bins



(d) $n = 12$ bins

Fig. 3: Histograms and frequency polygons, Example 3.5

- the value 139 stands alone suggesting that it is in fact an outlier;
- a Gamma family of distributions seems appropriate for CPU times, see the dashed curve in Figure 4;
- there is no indication of heterogeneity; all data points except $x = 139$ form a rather homogeneous group that fits the sketched Gamma curve.

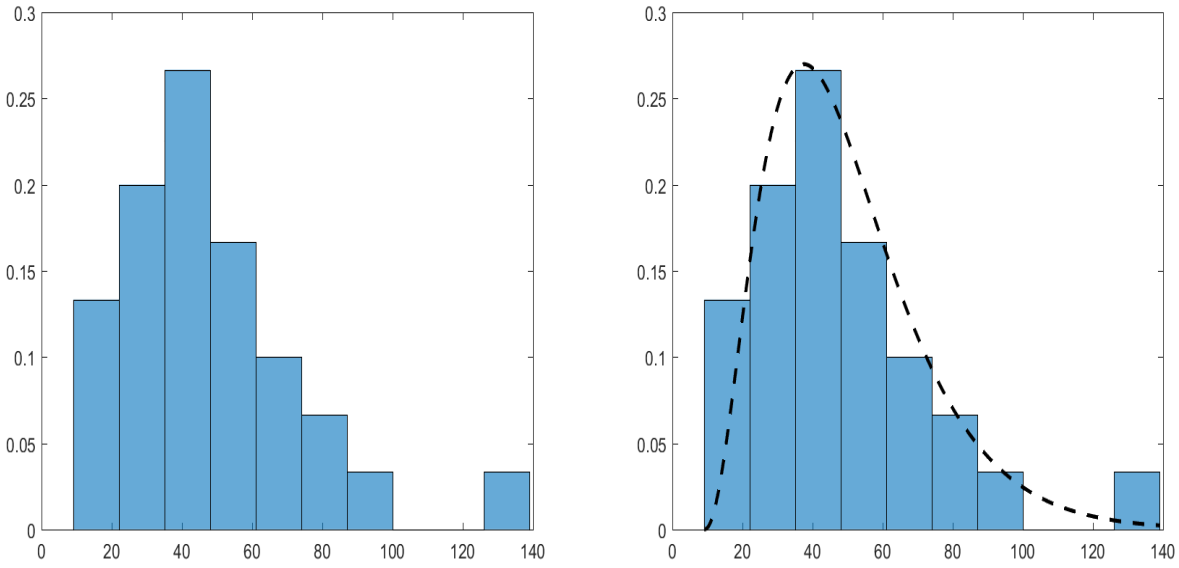


Fig. 4: Approximation of the pdf, Example 3.5

4 Calculative Descriptive Statistics

In the previous 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. Simple **descriptive statistics** measuring the location, spread, variability and other characteristics can be computed immediately. It is hoped that these will give us some idea of the values of the parameters that characterize the entire population from which the sample was pooled. 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.

4.1 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.

Mean

Definition 4.1. *The **(arithmetic) mean*** (mean) *of the data $x_1, \ldots, x_N$ is the value*

$$\overline{x}_a = \frac{1}{N} \sum_{i=1}^{N} x_i. \tag{4.2}$$

For grouped data, $\begin{pmatrix} x_i \\ f_i \end{pmatrix}_{i=\overline{1,n}}$, *the **mean** is* $\overline{x}_a = \frac{1}{N} \sum_{i=1}^{n} f_i x_i$.

Remark 4.2. 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 $a \in \mathbb{R}$,

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

Example 4.3. Let us recall the data in Example 3.5, where to evaluate the effectiveness of a processor, a sample of CPU times for $N = 30$ randomly chosen jobs (in seconds) was considered:

70	36	43	69	82	48	34	62	35	15
59	139	46	37	42	30	55	56	36	82
38	89	54	25	35	24	22	9	56	19

The *mean* CPU time is

$$\overline{x} \;=\; \frac{70 + 36 + \ldots + 56 + 19}{30} \;=\; 48.2333 \text{ seconds.}$$

We may conclude that the mean CPU time of *all* the jobs handled by that particular processor is about the same, "near" 48.2333 seconds. In other words, we try to estimate the *population mean* by the *sample mean*. How good would that approximation be? We will learn later how to assess the accuracy of our estimates.

Example 4.4. Let us assume that the value $x = 139$ (that seemed extreme, out of place, when we looked at the histogram) was *not* in this sample. Then the mean would be

$$\overline{x}_1 \ = \ 45.1034,$$

somewhat lower.
Now, in the other direction, let us suppose that the CPU time of one more job (a heavier one) is recorded and it is found to be 30 minutes $= 1800$ seconds. The mean of the new sample is

$$\overline{x}_2 \ = \ 104.7419 \text{ seconds},$$

way larger than the first value!

Median

One disadvantage of the sample mean is its *sensitivity to extreme observations*. As we have seen in the previous example, one extreme value can significantly shift the value of the mean, to the point where it becomes almost irrelevant.
The next measure of location is the *median*, which is much less sensitive than the mean.

Definition 4.5. *The **median*** ($\boxed{\text{median}}$) *is the value $\overline{M}$ 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.*

A sample is always *discrete*, since it consists of a finite number of observations. Then, computing a sample median is similar to the case of discrete distributions. In simple random sampling, all observations are equally likely, and thus, equal probabilities on each side of a median translate into an equal number of observations. There are two cases, depending on the sample size N.

If the sorted primary data is

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

then

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

Remark 4.6. The median may or may not be one of the values in the data.

Example 4.7. Let us find the median for the data in Example 4.3 (the CPU times).

Since there are $N = 30$ observations, there are two middle values, the 15th and the 16th entries.

$$\begin{array}{cccccccccc} 9 & 15 & 19 & 22 & 24 & 25 & 30 & 34 & 35 & 35 \\ 36 & 36 & 37 & 38 & \mathbf{42} & \mathbf{43} & 46 & 48 & 54 & 55 \\ 56 & 56 & 59 & 62 & 69 & 70 & 82 & 82 & 89 & 139 \end{array}$$

Then the median is $\overline{M} = 42.5$.

Remark 4.8. For an even number of observations, the median can be chosen to be any number between the two middle values. So in the previous example, we could say that any number in the interval $(42, 43)$ is a median.

Example 4.9. Let us add again the extreme value of 30 minutes $= 1800$ seconds. The new sample

$$\begin{array}{ccccccccccc} 9 & 15 & 19 & 22 & 24 & 25 & 30 & 34 & 35 & 35 & \\ 36 & 36 & 37 & 38 & 42 & \mathbf{43} & 46 & 48 & 54 & 55 & \\ 56 & 56 & 59 & 62 & 69 & 70 & 82 & 82 & 89 & 139 & 1800 \end{array}$$

has 31 observations, there is only one middle value (the 16th entry), so the median of the new sample is

$$\overline{M}_2 \ = \ 43.$$

Notice that the new value differs very little from the previous one and is *still relevant*, unlike the mean. So the median is a *robust* statistic, not being influenced (so much) by outliers.

Mode

Definition 4.10. *A* ***sample mode****, $\overline{x}_{mo}$, of a set of data is a most frequent value.*

Remark 4.11. 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**.

For data drawn from symmetric distributions, we have

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

In general,

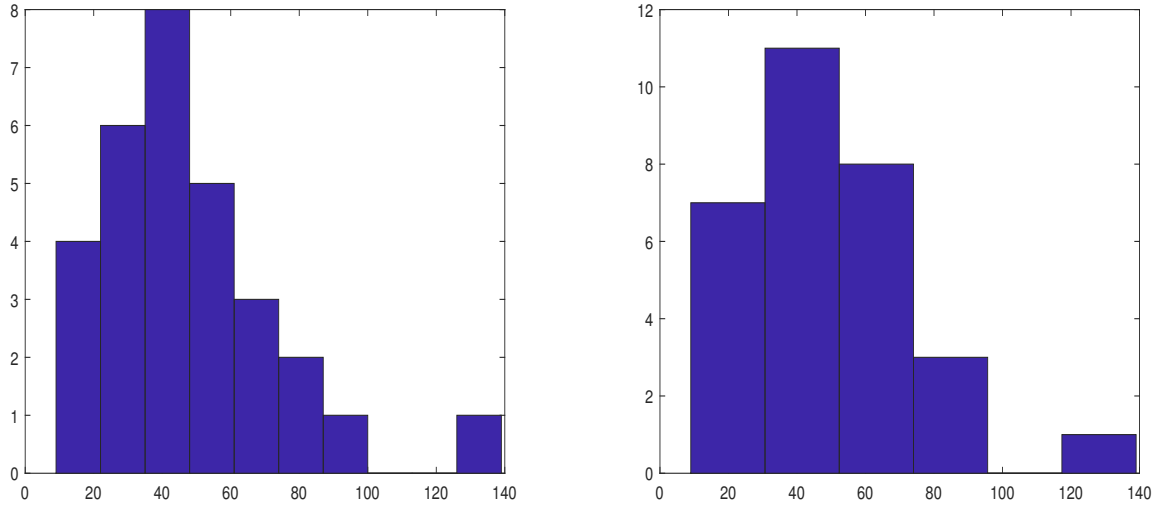
$$x_{mo} \approx \overline{x} - 3(\overline{x} - \overline{M}).$$

This empirical formula was given by K. Pearson.

Example 4.12. In our example about the CPU times, the values $35, 36, 56$ and 82 appear twice, while all the other values have a frequency of 1. So all four are modes, this is multimodal data.

9	15	19	22	24	25	30	34	**35**	**35**
36	**36**	37	38	42	43	46	48	54	55
56	**56**	59	62	69	70	**82**	**82**	89	139

If we group the data into 10 classes, then the *modal class* is the third one, $(35, 48]$, with modal mark 41.5 (Figure 5(a)). If we have only 6 classes, then the second one is the modal class, $[30.7, 52.4)$, with mark 41.55 (Figure 5(b)).



(a) $n = 10$ bins (b) $n = 6$ bins

Fig. 5: Modal class

4.2 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.

Quantiles, percentiles and quartiles

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 4.13. *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 4.14. *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*

$$MAD_2 = \text{median}\{|x_i - \overline{M}|\}.$$

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

(1) *A **sample p-quantile*** (quantile) *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*** (prctile) *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*

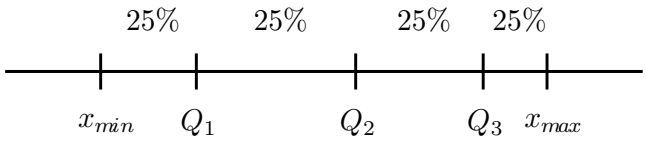


Fig. 6: Quartiles

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

Definition 4.16. *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*

$$IQR = Q_3 - Q_1. \tag{4.3}$$

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

$$IQD = \frac{IQR}{2} = \frac{Q_3 - Q_1}{2}. \tag{4.4}$$

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

$$IQDC = \frac{IQD}{\overline{M}} = \frac{Q_3 - Q_1}{2Q_2}. \tag{4.5}$$

Remark 4.17.

1. The interquartile deviation is an absolute measure of variation and it has an important property: the range $\overline{M} \pm IQD$ 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.

Example 4.18. Recall our example about the CPU times (in seconds) for $N = 30$ randomly chosen jobs (sorted ascendingly):

9	15	19	22	24	25	30	**34**	35	35
36	36	37	38	42	43	46	48	54	55
56	56	**59**	62	69	70	82	82	89	139

Let us compute various measures of variation.

Solution. For this example, the range is

$$139 - 9 = 130 \text{ seconds}$$

and the mean and median absolute deviations are

$$\begin{aligned} MAD_1 &= 19.6133, \\ MAD_2 &= 13.5. \end{aligned}$$

To determine the quartiles, notice that 25% of the sample equals $30/4 = 7.5$ and 75% of the sample is $90/4 = 22.5$ observations. From the ordered sample, we see that the 8th element, 34, has 7 observations to its left and 22 to its right, so it has *no more* than 7.5 observations to the left and *no more* than 22.5 observations to the right of it. Hence, $Q_1 = 34$.

Similarly, the third quartile is the 23rd smallest element, $Q_3 = 59$. Recall from last time that the second quartile (the median) is $Q_2 = \overline{M} = 42.5$. Then

$$\begin{aligned} IQR &= 59 - 34 = 25, \\ IQD &= IQR/2 = 12.5, \\ IQDC &= IQD/Q_2 = 0.2941. \end{aligned}$$

The interval

$$\overline{M} \pm IQD = [30, 55]$$

contains 14 observations.

The value of the $IQDC$ is close neither to 0, nor to the values ± 1. So the data doesn't show strong symmetry or strong asymmetry. This may be due to the extreme values 9 and/or 139.

■

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.

We have seen how the mean is very sensitive to outliers. Other statistical procedures can be gravely affected by the presence of outliers in the data. 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"? First, we could use a simple property, known as the "3σ rule". This is an application of Chebyshev's inequality

$$P(|X - E(X)| < \varepsilon) \geq 1 - \frac{V(X)}{\varepsilon^2},\ \forall \varepsilon > 0.$$

If we use the classical notations $E(X) = \mu, V(X) = \sigma^2,\ \text{Std}(X) = \sigma$ for the mean, variance and standard deviation of X and take $\varepsilon = 3\sigma$, we get

$$P(|X - \mu| < 3\sigma) \ \geq \ 1 - \frac{\sigma^2}{9\sigma^2} \ = \ \frac{8}{9} \ \approx \ .89.$$

This is saying that it is *very* probable (at least 0.89 probable) that $|X - \mu| < 3\sigma$, or, equivalently, that $\mu - 3\sigma < X < \mu + 3\sigma$. In words, the 3σ rule states that *most of the values that any random variable takes, at least* 89%*, lie within* 3 *standard deviations away from the mean*. This property is true in general, for any distribution, but especially for unimodal and symmetrical ones, where that percentage is even higher.

Based on that, 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.

A more general approach, that works well also for skewed data, is to consider an outlier any observation that is outside the range

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

Also, the coefficient $3/2$ can be replaced by some other number to decrease or enlarge the interval of "normal" values (or, equivalently, the domain that covers the outliers):

$$\left[Q_1 - w \cdot IQR,\ Q_3 + w \cdot IQR\right],\ w = 0.5, 1, 1.5.$$

For our example on CPU times of processors, we have

$$Q_1 - \frac{3}{2}IQR \ = \ -3.5,$$

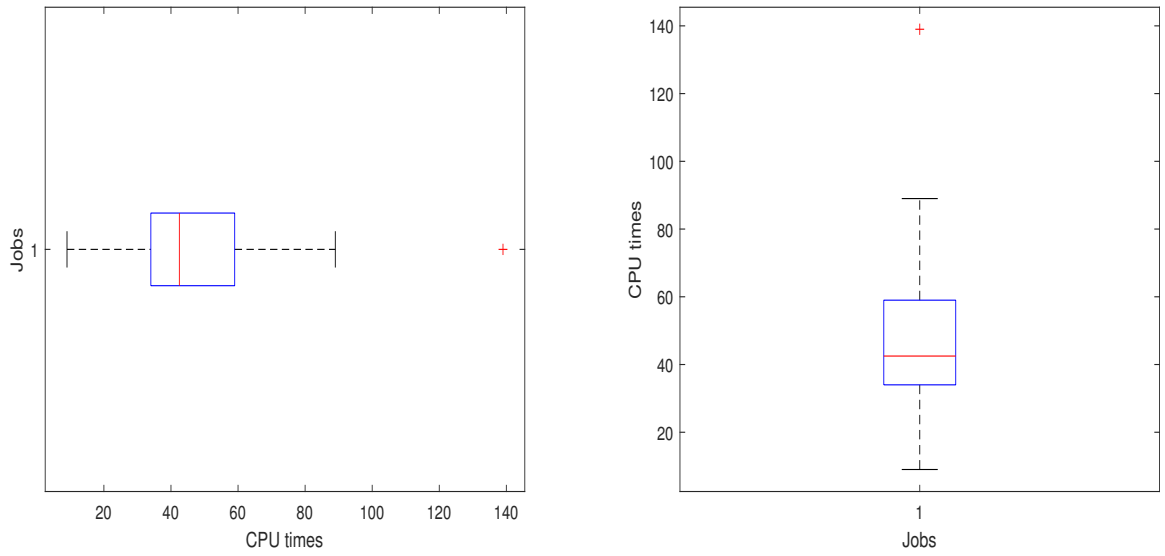
$$Q_3 + \frac{3}{2}IQR \ = \ 96.5,$$

so observations outside the interval $[-3.5, 96.5]$ are considered outliers. In this case, there is only one, the value 139.

Boxplots

All the information we discussed above is summarized in a graphical display, called a **boxplot** (boxplot), a plot in which a rectangle is drawn to represent the second and third quartiles (so the interquartile range), with a line inside for the median value and which indicates which values are considered extreme. The "whiskers" of the boxplot are the endpoints of the interval on which normal values lie (so everything outside the whiskers is considered an outlier).

For the data in Example 4.18, the boxplot is displayed in Figure 7 and it can be drawn vertically (default) or horizontally. The width of the interval of the whiskers can be changed. The interval that



(a) horizontally (b) vertically

Fig. 7: Boxplots

determines the outliers (i.e., outside of which values are considered too extreme, outliers) is

$$[Q_1 - w \cdot IQR, Q_3 + w \cdot IQR].$$

The default value is $w = 1.5$. With the smaller whiskers, boxplot displays more data points as outliers.

Moments, variance, standard deviation and coefficient of variation

The idea of the mean can be generalized, by taking various powers of the values in the data.

Definition 4.19.

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

$$\overline{\nu}_k = \frac{1}{N}\sum_{i=1}^{N} x_i^k, \;\; \overline{\nu}_k = \frac{1}{N}\sum_{i=1}^{n} f_i x_i^k, \tag{4.6}$$

for primary and for grouped data, respectively.

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

$$\overline{\mu}_k = \frac{1}{N}\sum_{i=1}^{N} (x_i - \overline{x})^k, \;\; \overline{\mu}_k = \frac{1}{N}\sum_{i=1}^{n} f_i (x_i - \overline{x})^k \tag{4.7}$$

for primary and for grouped data, respectively.

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

$$\overline{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N} (x_i - \overline{x})^2, \;\; \overline{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{n} f_i (x_i - \overline{x})^2 \tag{4.8}$$

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

Remark 4.20.

1. A more efficient computational formula for the variance is

$$\overline{\sigma}^2 = \frac{1}{N}\left(\sum_{i=1}^{N} x_i^2 - \frac{1}{N}\Big(\sum_{i=1}^{N} x_i\Big)^2\right) = \frac{1}{N}\left(\sum_{i=1}^{N} x_i^2 - N\overline{x}^2\right), \tag{4.9}$$

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 is

$$\begin{aligned} s^2 &= \frac{1}{N-1}\sum_{i=1}^{N} (x_i - \overline{x})^2 &&= \frac{1}{N-1}\Big(\sum_{i=1}^{N} x_i^2 - N\overline{x}^2\Big), \\ s^2 &= \frac{1}{N-1}\sum_{i=1}^{n} f_i (x_i - \overline{x})^2 &&= \frac{1}{N-1}\Big(\sum_{i=1}^{N} f_i x_i^2 - N\overline{x}^2\Big), \end{aligned} \tag{4.10}$$

for the *sample* variance for primary or grouped data. The reason the sum is divided by $N-1$ involves the notion of *degrees of freedom*, which takes into account the number of constraints in computing an estimate. The sample observations $x_1, \ldots, x_N$ are independent (by the definition of a random sample), but when computing the variance, we use the variables $x_1 - \overline{x}, \ldots, x_N - \overline{x}$. Notice that by subtracting the sample mean $\overline{x}$ from each observation, there exists a linear relation among the elements, namely

$$\sum_{k=1}^{N} (x_k - \overline{x}) = 0$$

and, thus, we lose 1 degree of freedom due to this constraint. Hence, there are only $N-1$ degrees of freedom. So, we will use (4.9) to compute the variance of a set of data that represents a population and (4.10) for the variance of a sample.

Example 4.21. Consider again our previous example on CPU times (in seconds) for $N = 30$ randomly chosen jobs:

70	36	43	69	82	48	34	62	35	15
59	139	46	37	42	30	55	56	36	82
38	89	54	25	35	24	22	9	56	19

Recall that for this data the sample mean was $\overline{x} = 48.2333$ seconds. The sample variance is

$$s^2 = \frac{(70 - 48.2333)^2 + \ldots + (19 - 48.2333)^2}{30 - 1} = \frac{20391}{29} \approx 703.1506 \text{ sec}^2.$$

Alternatively, using (4.9),

$$s^2 = \frac{70^2 + \ldots + 19^2 - 30 \cdot 48.2333^2}{30 - 1} = \frac{90185 - 69794}{29} \approx 703.1506 \text{ sec}^2.$$

The sample standard deviation is

$$s = \sqrt{703.1506} \approx 26.1506 \text{ sec}.$$

By the 3σ rule, using $\overline{x}$ and s as estimates for the population mean μ and population standard deviation σ, we may infer that at least 89% of the tasks performed by this processor require between $\overline{x} - 3s = -30.2185$ and $\overline{x} + 3s = 126.6851$ (so less than 126.6851) seconds of CPU time.

Definition 4.22. *The **coefficient of variation** is the value*

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

Remark 4.23.
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.

5 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 *response* and one or more independent variables, called *predictor(s)*. 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. Establishing and testing such a relation enables us: to understand interactions, causes, and effects among variables; to predict unobserved variables based on the observed ones; to determine which variables significantly affect the variable of interest, etc.

Example 5.1 (World Population)**.** According to the International Data Base of the U.S. Census Bureau, population of the world grows according to Table 4. How can we use these data to predict the world population in years 2025 and 2030?

Figure 8 shows that the population (response) is tightly related to the year (predictor). It increases every year, and its growth is almost linear. If we estimate the regression function relating our response and our predictor (see the dotted line on Figure 8) and extend its graph to the year 2030, the forecast is ready.

A straight line that fits the observed data for years $1950 - 2020$ predicts the population of 8.06 billion in 2025 and 8.444 billion in 2030. It also shows that between 2020 and 2025, the world

Year	Pop. (mln. people)	Year	Pop.(mln.people)	Year	Pop.(mln.people)
1950	2558	1975	4089	2000	6090
1955	2782	1980	4451	2005	6474
1960	3043	1985	4855	2010	6970
1965	3350	1990	5287	2015	7405
1970	3712	1995	5700	2020	7821

Table 4: World Population 1950-2020

population reaches the historical mark of 8 billion (which actually happened last year ...). How accurate is the forecast obtained in this example? The observed population during $1950 - 2020$ appears rather close to the estimated regression line in Figure 8. It is reasonable to hope that it will continue to do so through 2030.

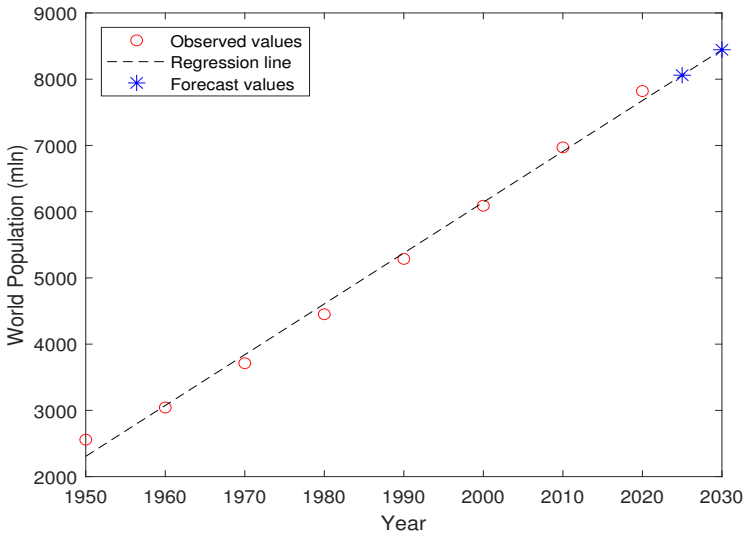


Fig. 8: World population and regression forecast

5.1 Univariate Regression, Curves of Regression

We will restrict our discussion to the case of **univariate regression**, predicting response Y based on *one* predictor X.

So, we have two vectors X and Y of 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 5.2. *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*

$$\overline{\nu}_{k_1 k_2} = \frac{1}{N} \sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} x_i^{k_1} y_j^{k_2}. \tag{5.1}$$

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

$$\overline{\mu}_{k_1 k_2} = \frac{1}{N} \sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} (x_i - \overline{x})^{k_1} (y_j - \overline{y})^{k_2}, \tag{5.2}$$

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

Remark 5.3. 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 5.4. *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*

$$\operatorname{cov}(X, Y) = \overline{\mu}_{11} = \frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}(x_i - \overline{x})(y_j - \overline{y}). \tag{5.3}$$

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

$$\overline{\rho} = \overline{\rho}_{XY} = \frac{\operatorname{cov}(X, Y)}{\sqrt{\overline{\mu}_{20}}\sqrt{\overline{\mu}_{02}}} = \frac{\overline{\mu}_{11}}{\overline{\sigma}_X \overline{\sigma}_Y}. \tag{5.4}$$

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 $\operatorname{cov}(X, Y) = \overline{\nu}_{11} - \overline{x} \cdot \overline{y}$.

The correlation coefficient is then

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

and, as before, it satisfies the inequality

$$-1 \leq \overline{\rho} \leq 1 \tag{5.5}$$

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 5.5. *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*

$$\overline{y}_i = \overline{y}(x_i) = \frac{1}{f_{i\cdot}} \sum_{j=1}^{n} f_{ij} y_j,\ i = \overline{1, m}. \tag{5.6}$$

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

$$\overline{x}_j = \overline{x}(y_j) = \frac{1}{f_{\cdot j}} \sum_{i=1}^{m} f_{ij} x_i,\ j = \overline{1, n}. \tag{5.7}$$

Definition 5.6. *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 5.7. 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).

5.2 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.

$$-2\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}\Big(y_j - f(x_i; a_1, \ldots, a_s)\Big)\frac{\partial f(x_i; a_1, \ldots, a_s)}{\partial a_k} = 0, \tag{5.8}$$

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 (5.8) becomes

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

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

$$y-\overline{y}=\overline{\rho}\,\frac{\overline{\sigma}_Y}{\overline{\sigma}_X}\,(x-\overline{x}) \tag{5.9}$$

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

$$x-\overline{x}=\overline{\rho}\,\frac{\overline{\sigma}_X}{\overline{\sigma}_Y}\,(y-\overline{y})\,. \tag{5.10}$$

Example 5.8. Let us consider the world population data in Example 5.1 and find the equation of the line of regression.

Solution. For the world population $(1950-2020)$ data, we find

$$\begin{aligned}\overline{x} &= 1985,\ \overline{y} = 4991.5\\ \overline{\sigma}_X &= 24.5,\ \overline{\sigma}_Y = 1884.6\\ \overline{\rho} &= 0.9972\end{aligned}$$

and the equation of the line of regression

$$y = 76.72x - 147300.5.$$

With this, we were able to forecast the values of 8.0604 billion for the year 2025 and 8.444 billion for 2030. Also, based on this model, the predicted population for 2024 is 7.9808 billion people.

■

Let us analyze linear regression further.

Remark 5.9.

1. The point of intersection of the two lines of regression (5.9) and (5.10) is $(\overline{x},\overline{y})$. This 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 we have the relation

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

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.

Example 5.10. Let us examine the situations graphed in Figure 9.

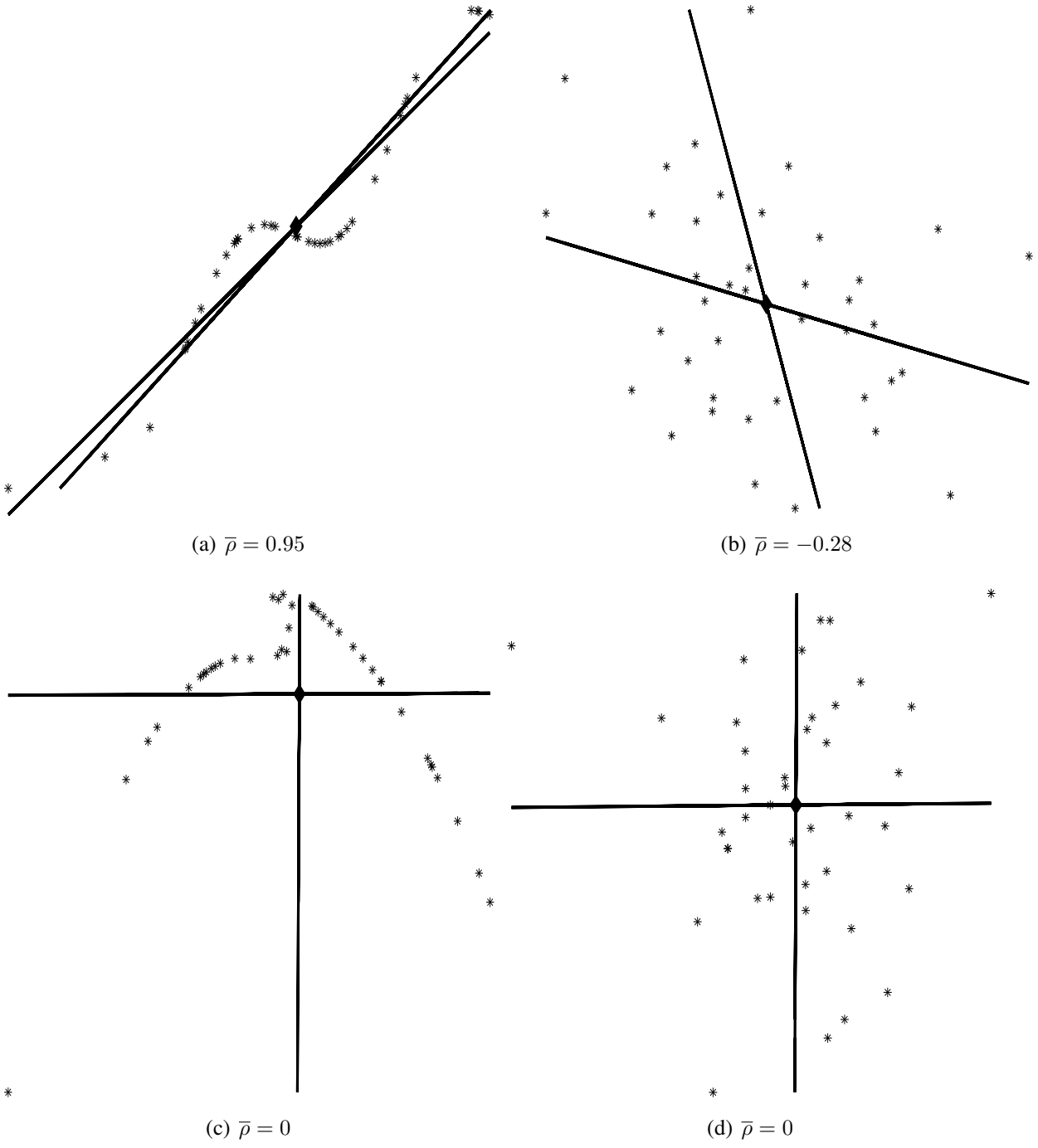
- In Figure 9(a) $\overline{\rho} = 0.95$, positive and very close to 1, suggesting a strong positive linear trend. Indeed, most of the points are on or very close to the line of regression of Y on X. The positivity indicates that large values of X are associated with large values of Y. Also, since the correlation coefficient is so close to 1, the two lines of regression almost coincide.

- In Figure 9(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 9(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 9(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 5.11. 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 $y = \dfrac{a}{x} + b$.



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

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

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

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

Fig. 9: Scattergram, Lines of Regression and Centroid