3.4 Quantiles

Let X be a random variable with cdf $F : \mathbb{R} \to \mathbb{R}$ and $\alpha \in (0, 1)$. A **quantile of order $\boldsymbol{\alpha}$** is a number q_α satisfying the condition $P(X < q_\alpha) \ \leq \ \alpha \ \leq \ P(X \leq q_\alpha)$, or, equivalently,

$$F(q_\alpha - 0) \ \leq \ \alpha \ \leq \ F(q_\alpha). \tag{3.1}$$

If X is continuous, then for each $\alpha \in (0, 1)$, there is a unique quantile q_α (see Figure 1), given by

$$F(q_\alpha) = \alpha,$$

or equivalently,

$$q_\alpha = F^{-1}(\alpha).$$

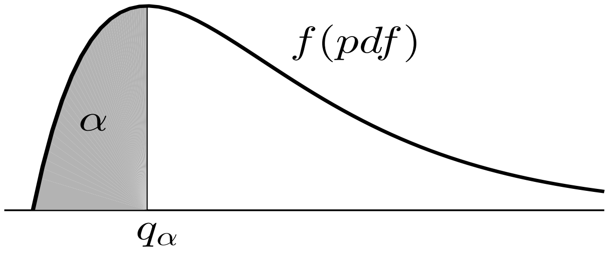


Fig. 1: Quantiles

3.5 Covariance and Correlation Coefficient

So far we have discussed numerical characteristics associated with one random variable. But oftentimes it is important to know if there is some kind of relationship between two (or more) random variables. So we need to define numerical characteristics that somehow measure that relationship.

Definition 3.1. *Let X and Y be random variables. The* ***covariance*** *of X and Y is the number*

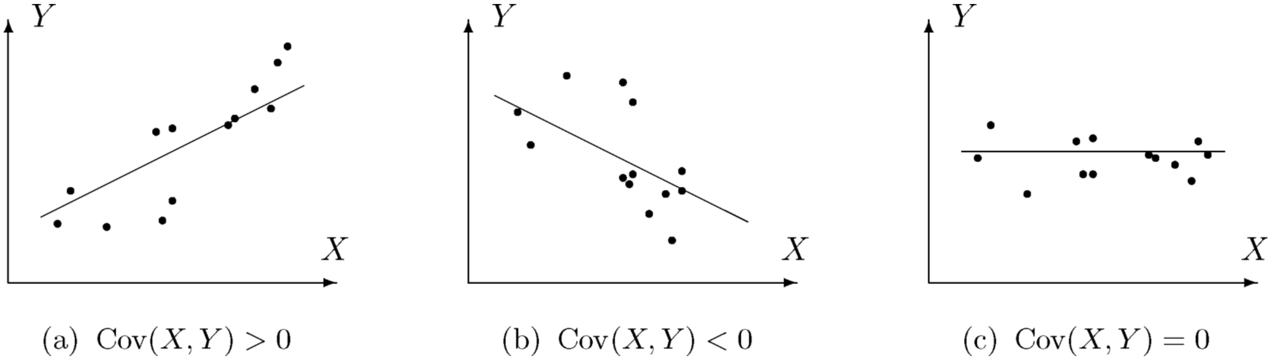
$$\operatorname{cov}(X, Y) \ = \ E\Big((X - E(X)) \cdot (Y - E(Y))\Big), \tag{3.2}$$

*if it exists. The **correlation coefficient** of X and Y is the number*

$$\rho(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sqrt{V(X)V(Y)}} = \frac{\operatorname{cov}(X,Y)}{\sigma(X)\sigma(Y)}, \tag{3.3}$$

if $\operatorname{cov}(X,Y), V(X), V(Y)$ *exist and* $V(X) \neq 0, V(Y) \neq 0$.

Notice the similarity between the definition of the covariance and that of the variance. The covariance measures the variation of two random variables *with respect to each other*. Just like with variance, large values (in absolute value) of the covariance show a strong relationship between X and Y, while small absolute values suggest a weak relationship. Unlike variance, covariance can also be negative. A negative value means that as the values of one variable increase, the values of the other decrease (see Figure 2).



(a) $\operatorname{Cov}(X,Y) > 0$ (b) $\operatorname{Cov}(X,Y) < 0$ (c) $\operatorname{Cov}(X,Y) = 0$

Fig. 2: Covariance

The covariance and correlation coefficient have the following properties:

- $\operatorname{cov}(X,X) = V(X)$;
- $\operatorname{cov}(X,Y) = E(XY) - E(X)E(Y)$ (a more efficient computational formula);
- If X and Y are independent, then $\operatorname{cov}(X,Y) = \rho(X,Y) = 0$ (we say that X and Y are **uncorrelated**);
- $V(aX + bY) = a^2V(X) + b^2V(Y) + 2ab\operatorname{cov}(X,Y)$, for all $a, b \in \mathbb{R}$;
- $\operatorname{cov}(X + Y, Z) = \operatorname{cov}(X,Z) + \operatorname{cov}(Y,Z)$;
- $|\rho(X,Y)| \leq 1$, i.e. $-1 \leq \rho(X,Y) \leq 1$;

- $|\rho(X,Y)| = 1$ if and only if there exist $a, b \in \mathbb{R}$, $a \neq 0$, such that $Y = aX + b$.

The correlation coefficient $\rho(X,Y)$ measures the linear "trend" between the variables X and Y. When $\rho = \pm 1$, there is "perfect linear correlation", so all the points (X,Y) are on a straight line (see Figure 3). The closer its value is to ± 1, the "more linear" the relationship between X and Y is. This notion will be revisited in the next chapter.

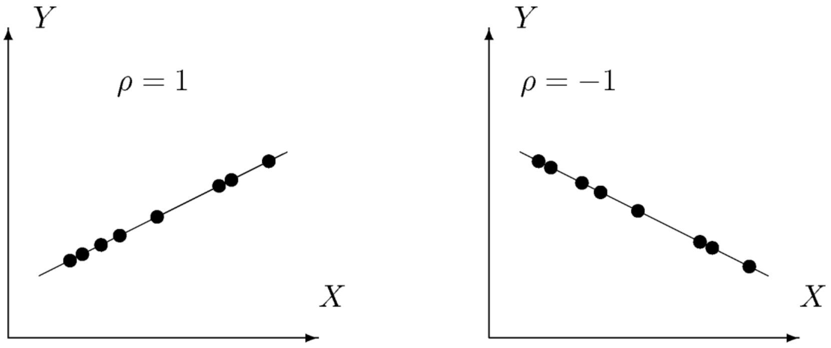


Fig. 3: Perfect correlation

Chapter 2. Introduction to Statistics

1 History of 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 scientific and more rigorous approach of Statistics was needed. The transition to inferential Statistics started at the beginning of last century, with the heavier employment of probabilistic methods.

As a discipline, Statistics has mostly developed in the past century. Probability theory - the mathematical foundation for Statistics - was developed in the $17\mathrm{th}$ to $19\mathrm{th}$ centuries based on work by Thomas Bayes, Pierre-Simon Laplace, and Carl Gauss. In contrast to the somewhat theoretical nature of probability, Statistics is an applied science concerned with analysis and modeling of data. Modern Statistics as a rigorous scientific discipline traces its roots back to the late 1800's and Francis Galton and Karl Pearson. R. A. Fisher, in the early $20\mathrm{th}$ century, was a leading pioneer of modern Statistics, introducing key ideas of *experimental design* and *maximum likelihood estimation*.

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 proposed a reformation of Statistics, where statistical inference is just one component of data analysis. 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. His conjecture was that a statistical model may or may not be used, but primarily EDA is for seeing what the data can tell us beyond the formal modeling and thereby somehow contrasts traditional hypothesis testing. The engineering and computer science communities quickly embraced this new approach of analyzing data sets. 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, access to more and bigger data, and the greater use of quantitative analysis in a variety of disciplines.

As a consequence, new disciplines in Statistics were established, such as *Robust statistics* and *Nonparametric tests*, which do not rely so heavily on theoretical assumptions and are not so easily affected by outliers.

2 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 (see Figure 4).
- 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);

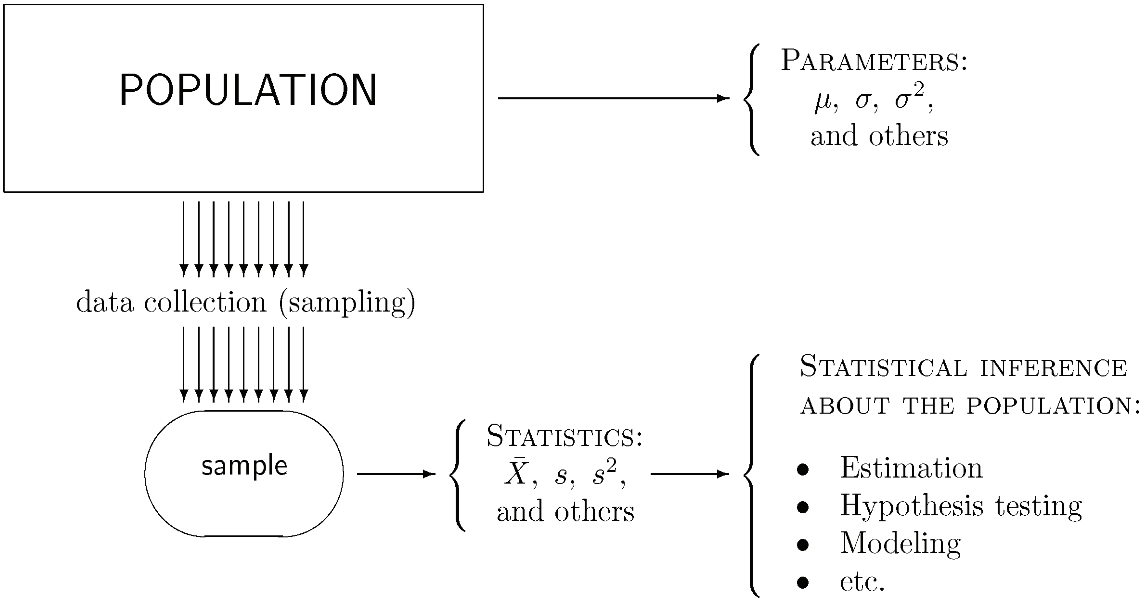


Fig. 4: Population parameters and sample statistics

  - **ordinal**, categorical data that has an explicit ordering.

- Data can also be classified as

  - **rectangular**, data in the form of a two-dimensional matrix with rows indicating records (cases) and columns indicating features (variables), like a table or spreadsheet; used in a **cross-sectional** study, which captures a snapshot of a group at a point in time;

  - **non-rectangular** data or time series; used in a **longitudinal** study, which observes a group repeatedly over a period of time.

- Any data analysis project has 5 important steps:

  1. Determining and declaring the objective(s).

  2. Collecting the data (Descriptive Statistics).

  3. Cleaning (wrangling) and organizing the data. Data wrangling is the process of removing errors and combining complex data sets to make them more accessible and easier to analyze (Descriptive Statistics).

  4. Analyzing the data (Inferential Statistics and Exploratory Data Analysis).

  5. Interpreting and sharing the results.

3 Data Collection

3.1 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

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.

Another important technique is **data mining**, which, in data science is defined as "the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data". Today's technologies have enabled the automated extraction of hidden predictive information from databases, along with a confluence of various other fields like artificial intelligence, machine learning, database management, pattern recognition, and data visualization. With data mining, an individual applies various methods of Statistics, data analysis and machine learning to explore and analyze large data sets, to extract new and useful information. Also, using data mining, an organization may discover actionable insights from their existing data.

3.2 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!

Example 3.1. A survey among passengers of some airline is conducted in the following way. A sample of random flights is selected from a list, and ten passengers on each of these flights are also randomly chosen. Each sampled passenger is asked to fill a questionnaire. Is this a representative sample? Suppose Mr. X flies only once a year, whereas Ms. Y has business trips twice a month. Obviously, Ms. Y has a *much higher* chance to be sampled than Mr. X. *Unequal* probabilities have to be taken into account, otherwise a non-sampling error will inevitably occur.

Example 3.2. (U. S. A. Presidential Election of 1936). A popular weekly magazine, *The Literary Digest*, correctly predicted the winners of 1920, 1924, 1928, and 1932 U. S. Presidential elections. However, it failed to do so in 1936! Based on a survey of ten million people, it predicted an overwhelming victory of Republican Governor Alfred Landon. Instead, the Democrat Franklin Delano Roosevelt received 98.49% of the electoral vote, won 46 out of 48 states, and was re-elected. This was the largest share of the Electoral College since 1820, the second-largest number of raw electoral votes ever received by a candidate, and the largest ever for a Democrat! So, what went wrong in that survey? At least two main issues with their sampling practice caused this prediction error. First, the sample was based on the population of subscribers of *The Literary Digest* that was dominated by Republicans. Second, responses were voluntary, and 77% of mailed questionnaires were not returned, introducing further bias.

4 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 4.1. 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 5 and bar graphs in Figure 6.

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**

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

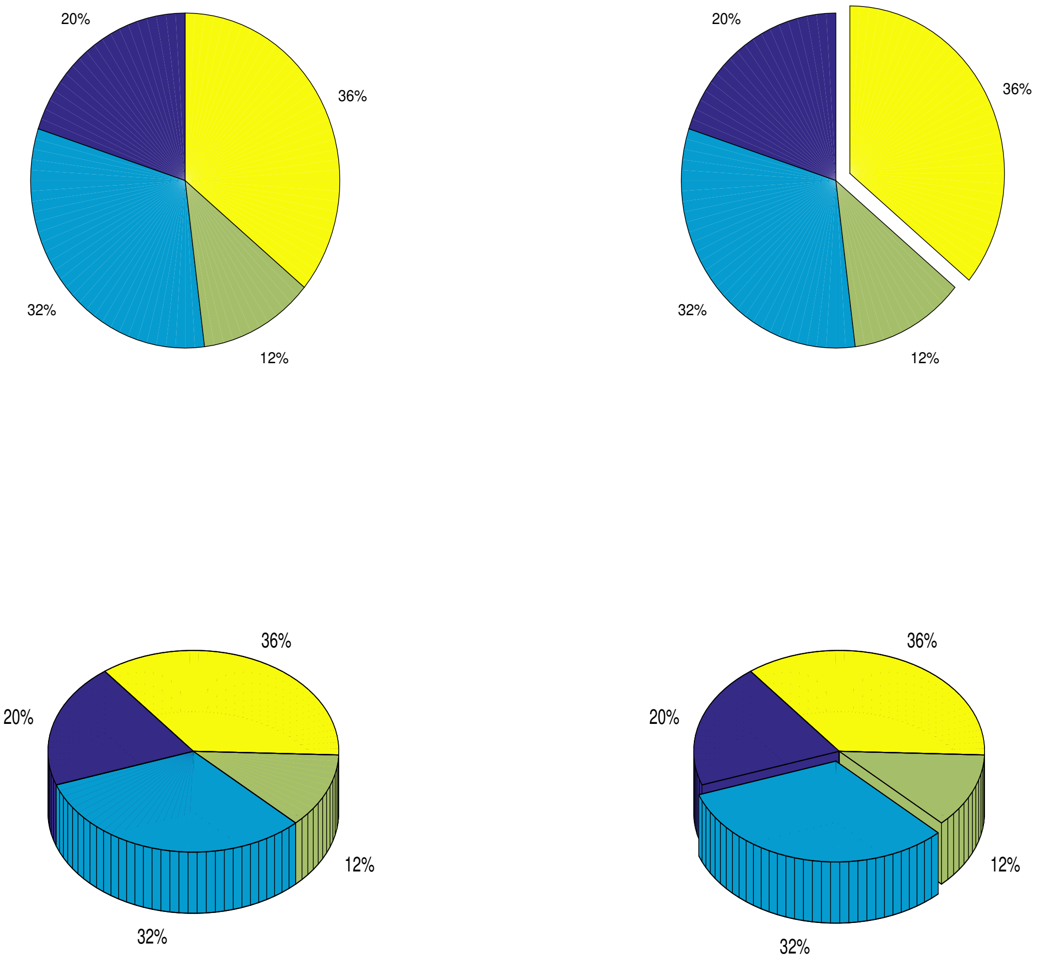


Fig. 5: Pie Charts

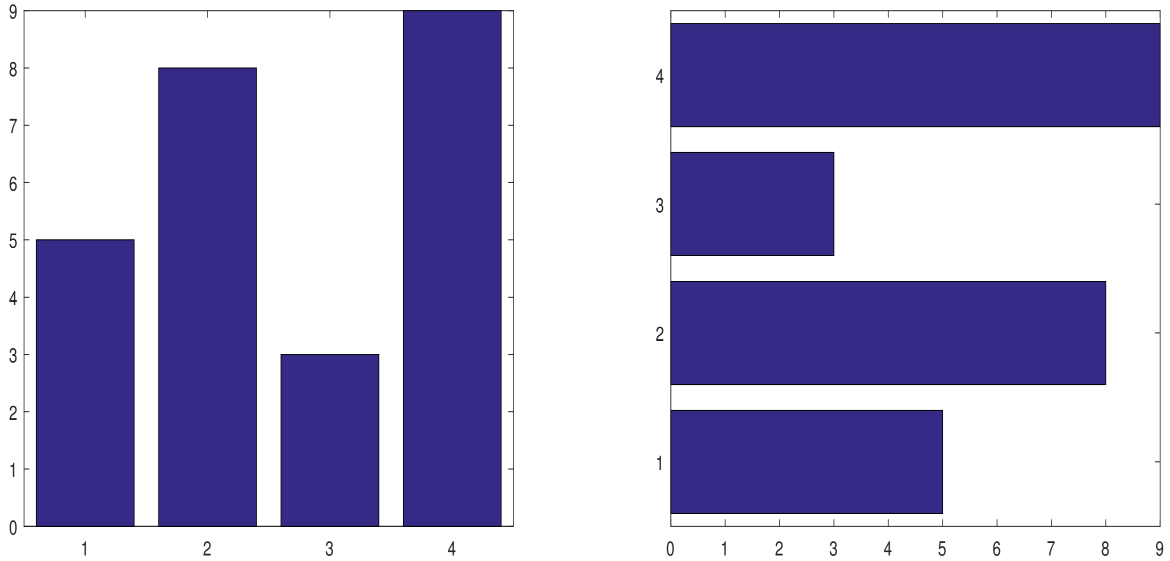


Fig. 6: Bar graphs

or **relative cumulative frequencies**

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

where $N = \sum\limits_{i=1}^{k} f_i$ 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** $x_i = \dfrac{c_{i-1}+c_i}{2}$, the midpoint of the interval, as an identifier for the class,

$-$ the **class width (length)** $l_i = c_i - c_{i-1}$,

$-$ the **class frequency** f_i, the sum of the frequencies of all observations x in that class.

Notice that we used the same notation x_i for primary data and for class marks. This is by choice, since in the case of grouped data, the class mark plays the role of a "representative" for that class and the class frequency is taken as being the frequency of that one value. The double notation should not cause confusion throughout the text, since N is the sample size, so $x_1, \ldots, x_N$ denotes the primary data, while n is the number of classes and thus,

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

denotes the grouped frequency distribution of the data.

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

Remark 4.2.
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*

$$n = 1 + \frac{10}{3}\log_{10}N, \tag{4.4}$$

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 4.3. 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 (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

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

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

There are $N = 30$ observations, with $x_{\min} = 9$ and $x_{\max} = 139$.
Since there are very few repetitions, the ungrouped frequency distribution table doesn't tell us much (see Table 2).
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 3 and 4. We have also included the relative and cumulative frequencies.

Remark 4.4. 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.

Histograms and Frequency Polygons

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

Value	Frequency
9	1
15	1
19	1
22	1
24	1
25	1
30	1
34	1
35	2
36	2
37	1
38	1
42	1
43	1
46	1
48	1
54	1
55	1
56	2
59	1
62	1
69	1
70	1
82	2
89	1
139	1

Table 2: Frequency Distribution Table for Example 4.3

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 3: Example 4.3, 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 4: Example 4.3, Grouped frequency distribution table with $n = 6$ classes

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.

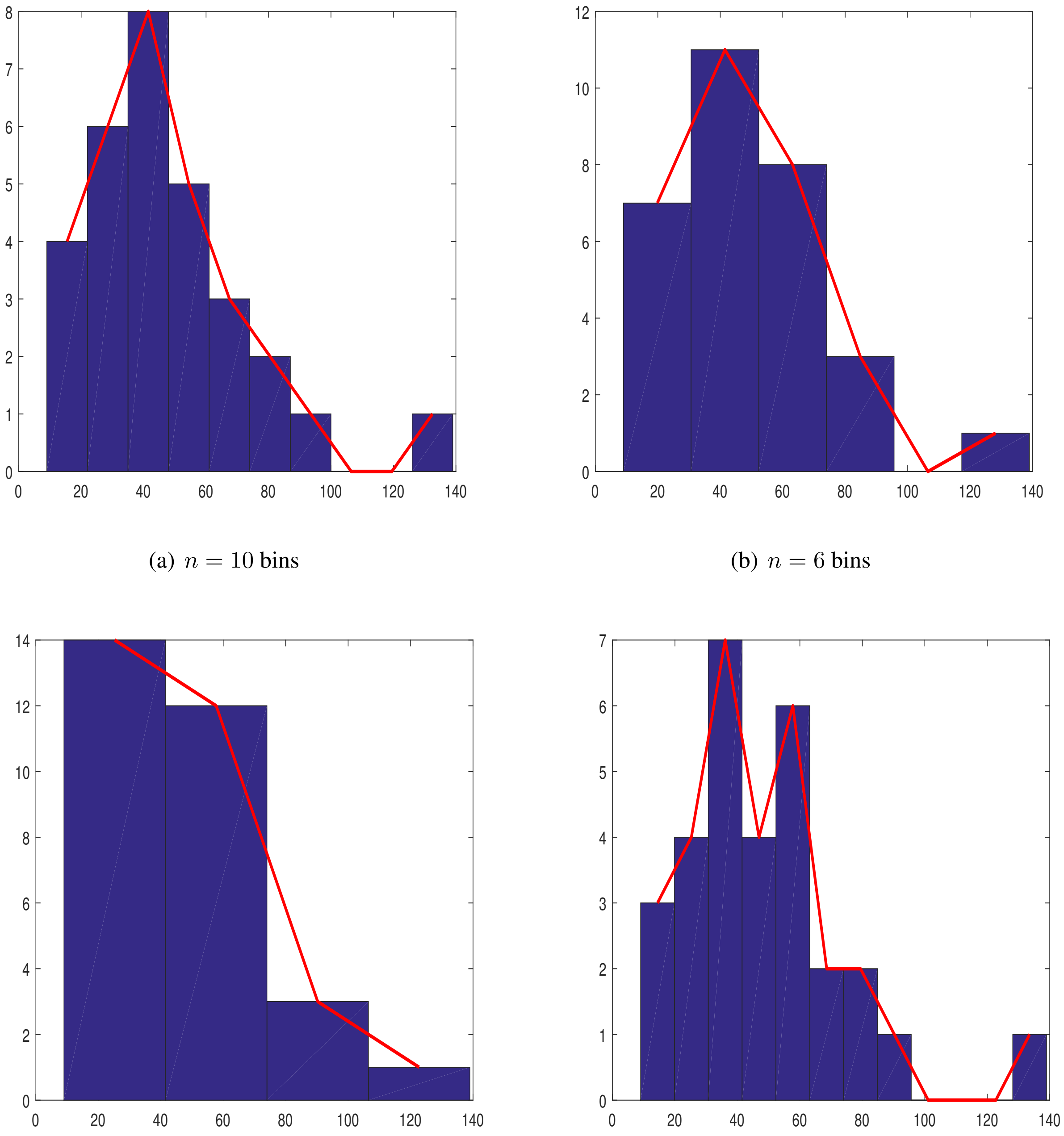
A **relative frequency histogram** has the same shape but a different vertical scale. Its column heights represent the proportion of all data that appeared in each bin.

If relative frequencies are considered (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 4.4. Consider the data from Example 4.3, the CPU times for $N = 30$ randomly chosen jobs.

We constructed the grouped frequency distribution tables for these data for $n = 10$ and for $n = 6$ classes. Figure 7 shows the corresponding histogram and frequency polygon for grouped data ((a) and (b)). Also in Fig. 7, 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 $n = 10$.



(a) $n = 10$ bins

(b) $n = 6$ bins

(c) $n = 4$ bins

(d) $n = 12$ bins

Fig. 7: Histograms and Frequency Polygons, Example 4.4

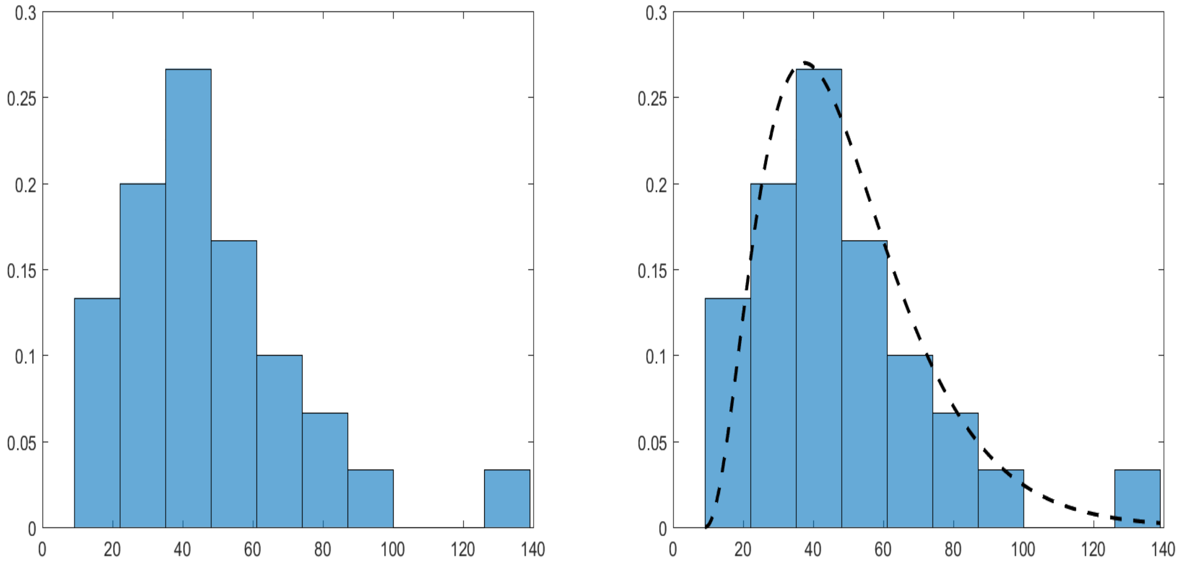


Fig. 8: Approximation of the pdf, Example 4.4

For 10 classes, let us take a closer look, see Figure 8. 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;
- 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 8;
- there is no indication of heterogeneity; all data points except $x = 139$ form a rather homogeneous group that fits the sketched Gamma curve.