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. The transition to inferential Statistics started at the beginning of last century, with the heavier employment of probabilistic methods. 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. 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 (extreme values).

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.

- 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.

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

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.

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!

Example 2.1. 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 2.2. (U. S. 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 Governor Alfred Landon. Instead, Franklin Delano Roosevelt received 98.49% of the electoral vote, won 46 out of 48 states, and was re-elected. 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.

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.

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**

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

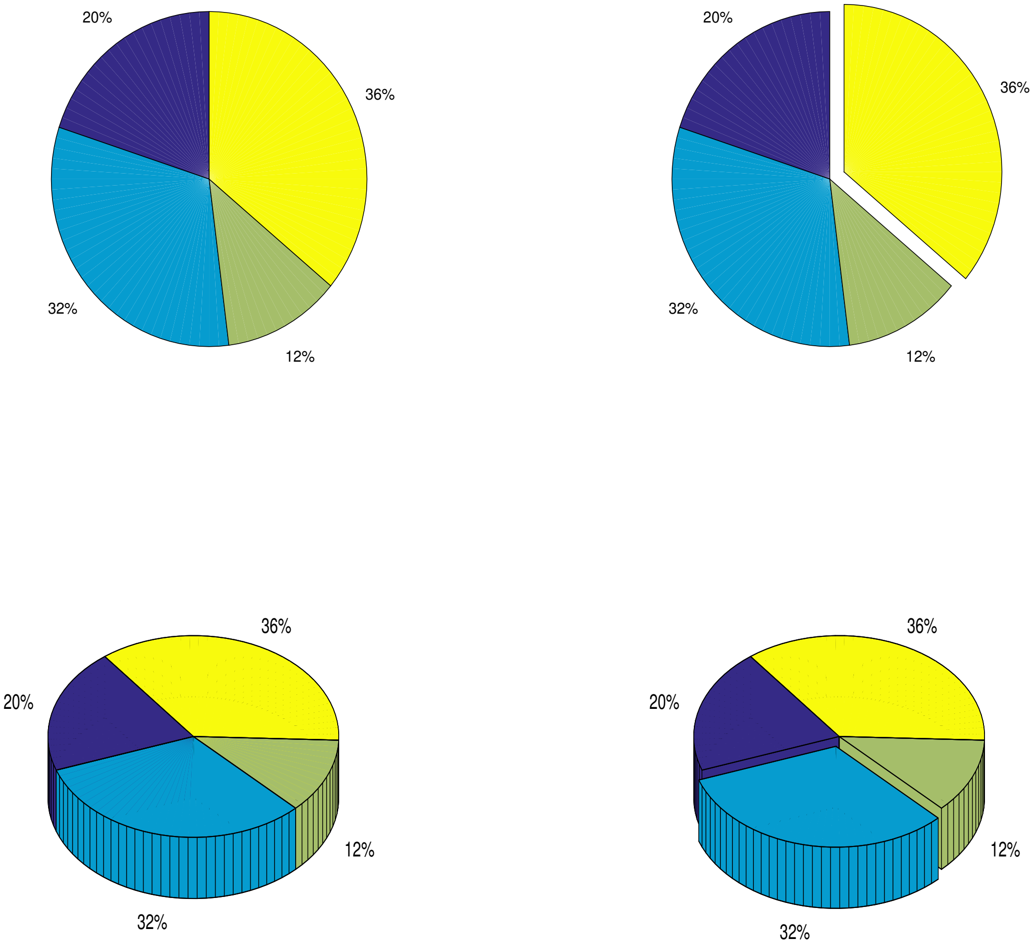


Fig. 1: Pie Charts

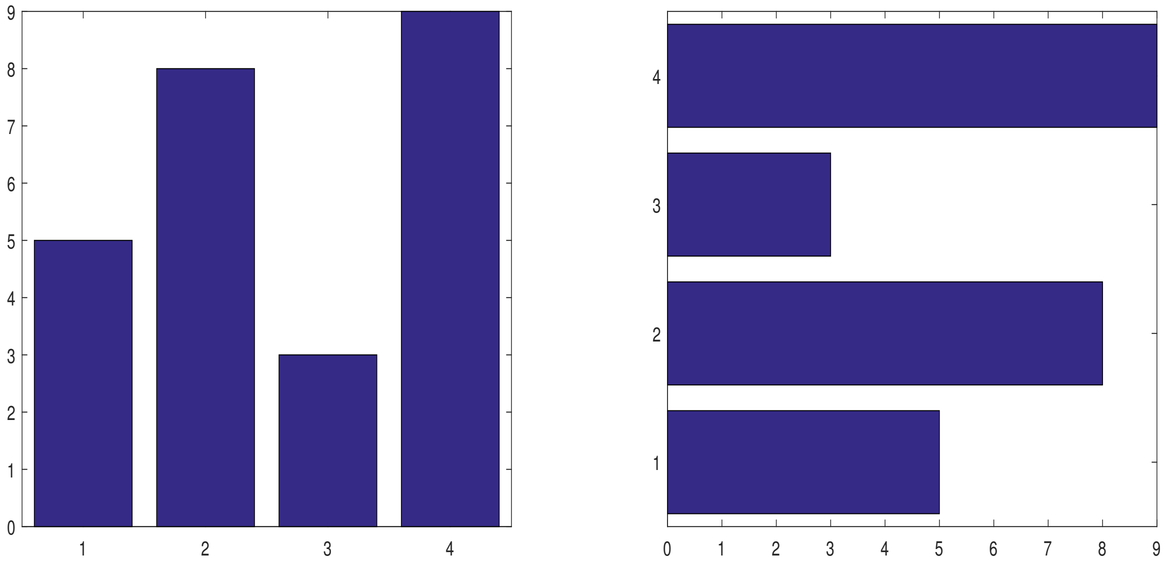


Fig. 2: Bar graphs

or **relative cumulative frequencies**

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

where $N = \sum_{i=1}^{k} f_i$ 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 $x_1, \dots, x_N$ 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:

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 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** ($\boxed{\text{hist/histogram}}$). 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 3.3

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 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 4: Example 3.3, 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 (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.

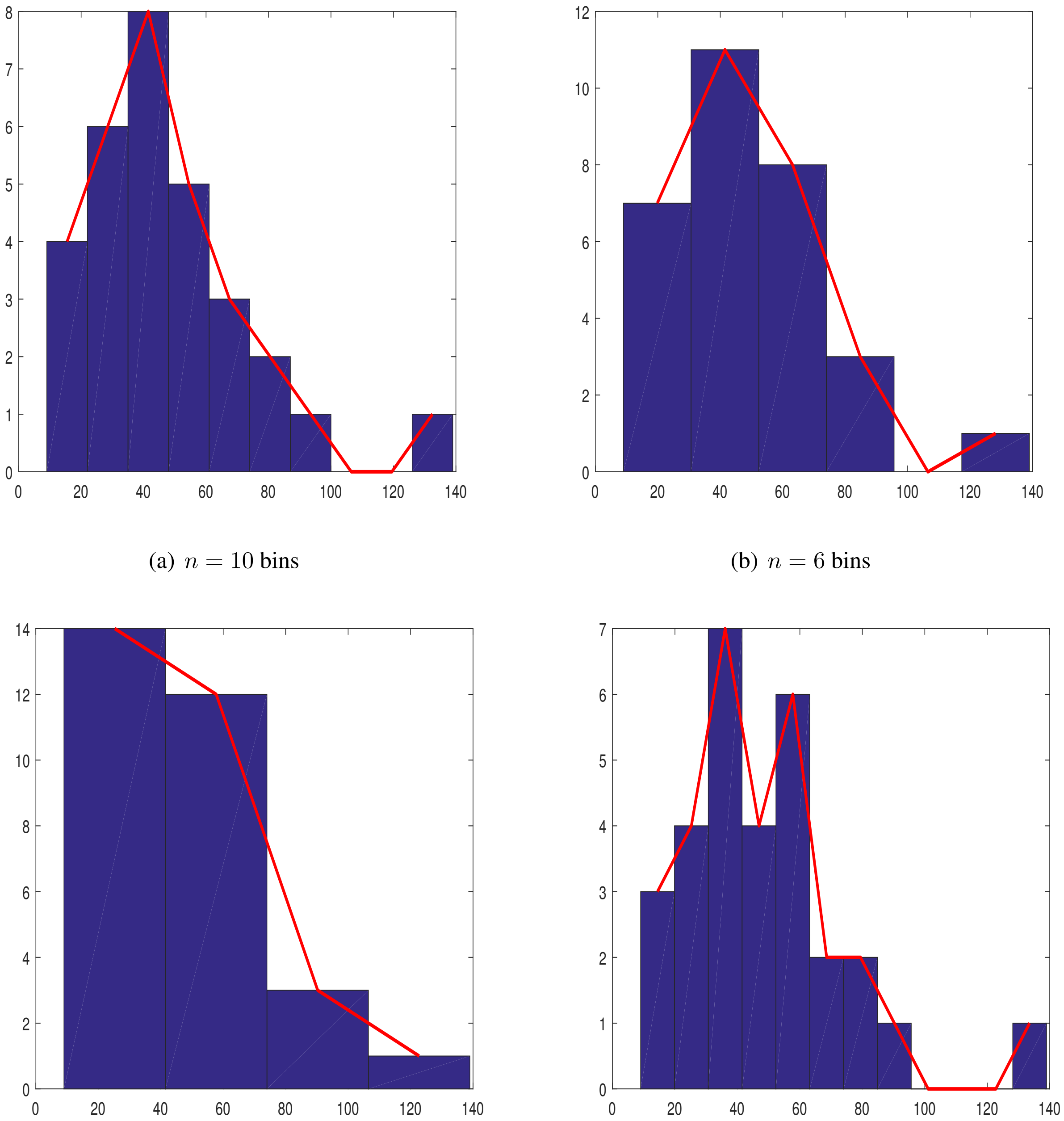
If relative frequencies are considered (so the proportionality factor is N, the total number of observations), then the total areas of all rectangles will be equal to 1. For a large volume of data grouped into a reasonably large number of classes, the histogram gives a rough approximation of the density 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 3.5. Let us consider again the data in Example 3.3, the CPU times (in seconds) for $N = 30$ randomly chosen jobs:

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

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 $n = 10$.



(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

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;
- 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.

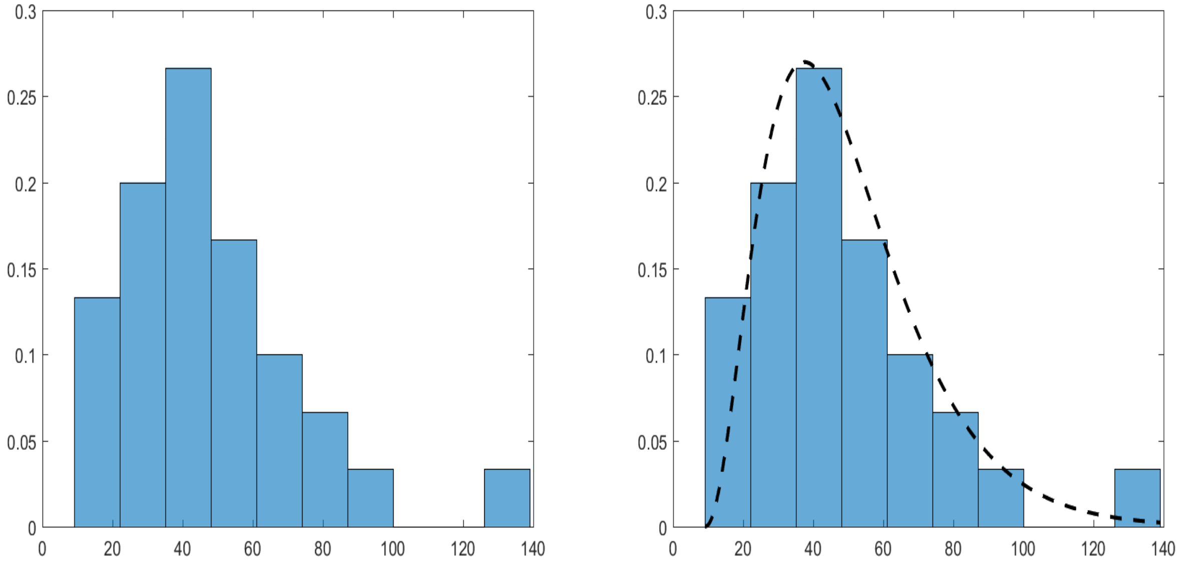


Fig. 4: Approximation of the pdf, Example 3.5

3.3 Stem-and-Leaf Plots

Stem-and-leaf plots are similar to histograms, although they carry more information. Namely, they also show how the data are distributed *within columns*. To construct a stem-and-leaf plot, we need to draw a stem and a leaf. The first one or several digits form a “stem”, and the next digit forms a "leaf". Other digits are dropped; in other words, the numbers get rounded. For example, the number 139 can be written as

$$\begin{array}{r|l} 13 & 9 \end{array}$$

with 13 going to the stem and 9 to the leaf, or as

$$\begin{array}{r|l} 1 & 3 \end{array}$$

with 1 joining the stem, 3 joining the leaf, and the digit 9 being dropped. In the first case, the leaf unit equals 1 and the stem unit is 10, while in the second case, the leaf unit is 10 and the stem unit is 10^2, showing that the (rounded) number is not 13, but 130. The stem and leaf units *must be carefully specified* for each such plot.

$$\begin{array}{r|llllllll}
0 & 9 \\
1 & 5 & 9 \\
2 & 2 & 4 & 5 \\
3 & 0 & 4 & 5 & 5 & 6 & 6 & 7 & 8 \\
4 & 2 & 3 & 6 & 8 \\
5 & 4 & 5 & 6 & 6 & 9 \\
6 & 2 & 9 \\
7 & 0 \\
8 & 2 & 2 & 9 \\
9 & \\
10 & \\
11 & \\
12 & \\
13 & 9
\end{array}$$

Example 3.6. For the CPU times in Example 3.5 (sorted increasingly),

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 draw a stem-and-leaf plot with leaf unit 1 (i.e., the last digits form a leaf). The remaining digits go to the stem, so the stem unit is 10. Each CPU time is then written as

$$10 \text{ “stem”} + \text{ “leaf”},$$

making the stem-and-leaf plot above.

Turning this plot by 90 degrees counterclockwise, we get a histogram with $10-$unit bins (because each stem unit equals 10). Thus, all the information seen on a histogram can be obtained here too. In addition, now we can see *individual* values within each column.

Stem-and-leaf plots can also be used to compare two samples. For this purpose, one can put two leaves on the same stem.

Example 3.7. The following two samples represent transmission times (in seconds) of signals - known as “pings”- from two different locations.

L1: 0.0156, 0.0396, 0.0355, 0.0480, 0.0419, 0.0335, 0.0543, 0.0350,
0.0280, 0.0210, 0.0308, 0.0327, 0.0215, 0.0437, 0.0483,
L2: 0.0298, 0.0674, 0.0387, 0.0787, 0.0467, 0.0712, 0.0045, 0.0167,
0.0661, 0.0109, 0.0198, 0.0039.

Let us sort the two samples in increasing order.

L1: 0.0156, 0.0210, 0.0215, 0.0280, 0.0308, 0.0327, 0.0335, 0.0350
0.0355, 0.0396, 0.0419, 0.0437, 0.0480, 0.0483, 0.0543,
L2: 0.0039, 0.0045, 0.0109, 0.0167, 0.0198, 0.0298, 0.0387, 0.0467
0.0661, 0.0674, 0.0712, 0.0787.

$$\begin{array}{r|c|l}
 & 0 & 3\ 4 \\
5 & 1 & 0\ 6\ 9 \\
1\ 1\ 8 & 2 & 9 \\
0\ 2\ 3\ 5\ 5\ 9 & 3 & 8 \\
1\ 3\ 8\ 8 & 4 & 6 \\
4 & 5 & \\
 & 6 & 6\ 7 \\
 & 7 & 1\ 8
\end{array}$$

Since all numbers start with 0.0..., we choose a stem unit of 0.01, a leaf unit of 0.001 and drop the last digit. We construct the above two stem-and-leaf plots (two in one), one to the left (L1) and one to the right (L2) of the stem. Looking at these two plots, we can see about the same average ping from the two locations. Also, we realize that the first location has a more stable connection, because its pings have lower variability (i.e., lower variance).

3.4 Scatter Plots and Time Plots

Scatter plots are used to see and understand a relationship between two variables. These can be temperature and humidity, experience and salary, age of a network and its speed, number of servers and the expected response time, etc. To study the relationship, both variables are measured on each sampled item. For example, temperature and humidity during each of n days, age and speed of n networks, or experience and salary of n randomly chosen employees are recorded. Then, a **scatter plot (scattergram)** consists of n points on an (x, y)-plane, with x- and y-coordinates representing the two recorded variables.

Example 3.8. Protection of a personal computer largely depends on the frequency of running antivirus software on it. One can set to run it every day, once a week, once a month, etc. During a scheduled maintenance of computer facilities, a computer manager records the number of times the antivirus software was launched on each computer during 1 month (variable X) and the number of detected viruses (worms) (variable Y). The data for 30 computers are given in the table below.

X	30	30	30	30	30	30	30	30	30	30	30	15	15	15	10
Y	0	0	1	0	0	0	1	1	0	0	0	0	1	1	0

X	10	10	6	6	5	5	5	4	4	4	4	4	1	1	1
Y	0	2	0	4	1	2	0	2	1	0	1	0	6	3	1

Is there a connection between the frequency of running antivirus software and the number of worms in the system? A scatter plot of these data is given in Figure 5(a). It clearly shows that the number of worms reduces, in general, when the antivirus is employed more frequently. This relationship, however, is not certain because no worm was detected on some "lucky" computers although the antivirus software was launched only once a week (4 times a month) on them.

Looking at the scatter plot in Figure 5(a), the manager realized that a portion of data is hidden there because there are identical observations. For example, no worms were detected on 8 computers where the antivirus software is used daily (30 times a month). Then, Figure 5(a) may be misleading. When the data contain identical pairs of observations, the points on a scatter plot are often depicted with either numbers or letters (e.g., "A" for 1 point, "B" for two identical points, "C" for three, ..., "H" for eight, etc.). You can see the result in Figure 5(b).

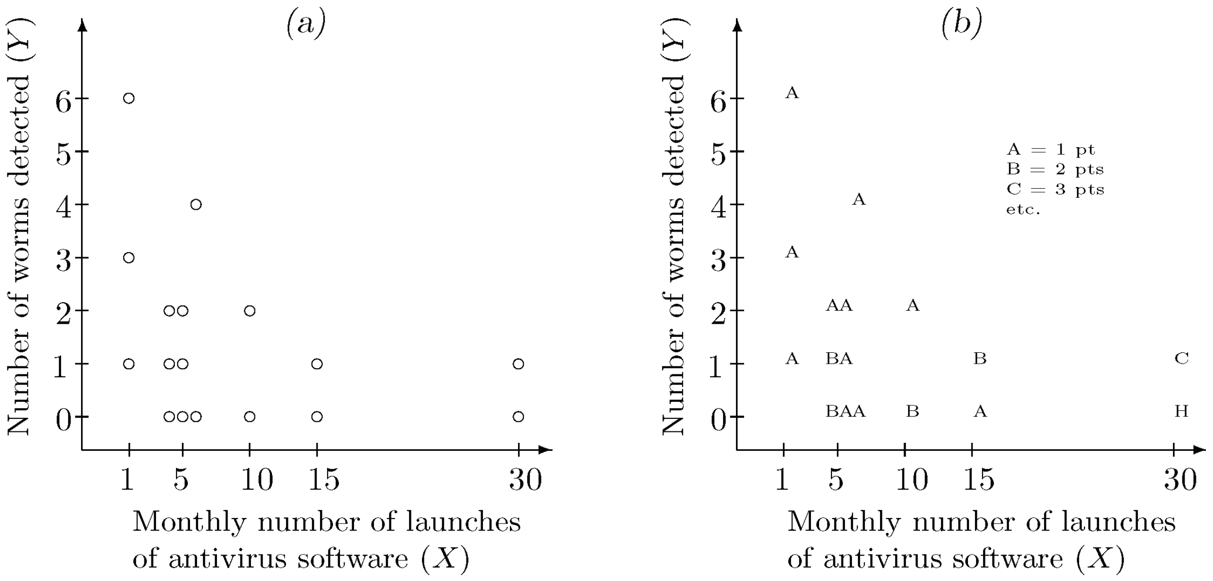


Fig. 5: Scatter plots for Example 3.8

When we study time trends and development of variables over time, we use **time plots**. These are scatter plots with x-variable representing time.

Example 3.9. Here is how the world population increased between 1950 and 2012 (Figure 6). We can clearly see that the population increases at an almost steady rate.

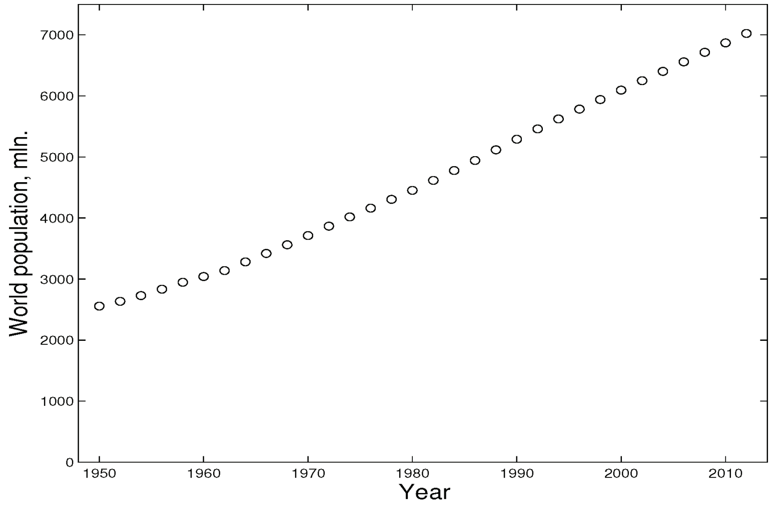


Fig. 6: Time plot of the world population in 1950–2012, Example 3.9

The actual data will be given and studied, later on (Correlation and Regression). We will estimate the trends seen on time plots and scatter plots and even make forecasts for the future.

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** ($\boxed{\text{mean}}$) of the data $x_1, \dots, x_N$ is the value*

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

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

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

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 $a \in \mathbb{R}$,

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

A straightforward computation leads to

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

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

Definition 4.3. *The* ***geometric mean*** ($\boxed{\text{geomean}}$) *of the data $x_1, \dots, x_N$ is the value*

$$\overline{x}_g = \sqrt[N]{x_1 \dots x_N}. \tag{4.3}$$

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 4.4. *The* ***harmonic mean*** ($\boxed{\text{harmmean}}$) *of the data $x_1, \dots, x_N$ is the value*

$$\overline{x}_h = \frac{N}{\displaystyle\sum_{i=1}^{N} \frac{1}{x_i}}. \tag{4.4}$$

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

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

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

Remark 4.5.
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 $\overline{x}$.

Example 4.6. 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

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

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 4.7. 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 4.8. *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 4.9. The median may or may not be one of the values in the data.

Example 4.10. Let us find the median for the data in Example 4.6 (the CPU times).

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

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

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

Remark 4.11. 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 4.12. Let us add again the extreme value of 30 minutes $= 1800$ seconds. The new sample

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

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 4.13. *A* ***mode*** Mo *of a random variable* X *is a value with the highest pdf, i.e., it is the point with the highest concentration of probability,* $Mo = \operatorname{argmax}\{f(x)\}$. *A* ***sample mode,*** $\overline{x}_{mo}$*, of a set of data is a most frequent value.*

Remark 4.14. 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

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

This is true, for instance, for the Normal distribution which is unimodal (Figure 7). For a Uniform $U(a, b)$ distribution, *all* values in the interval $[a, b]$ are modes (Figure 8), while the $\chi^2(1)$ distribution (with $\nu = 1$ degree of freedom) has no mode (Figure 9).

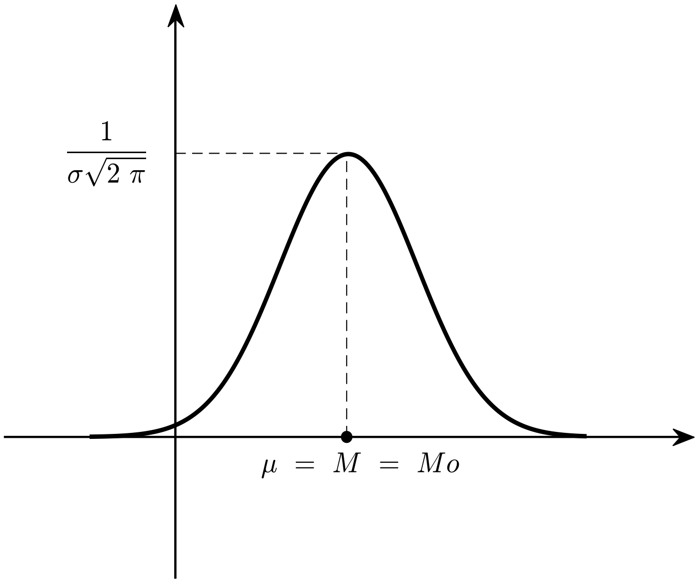


Fig. 7: Normal Distribution (unimodal)

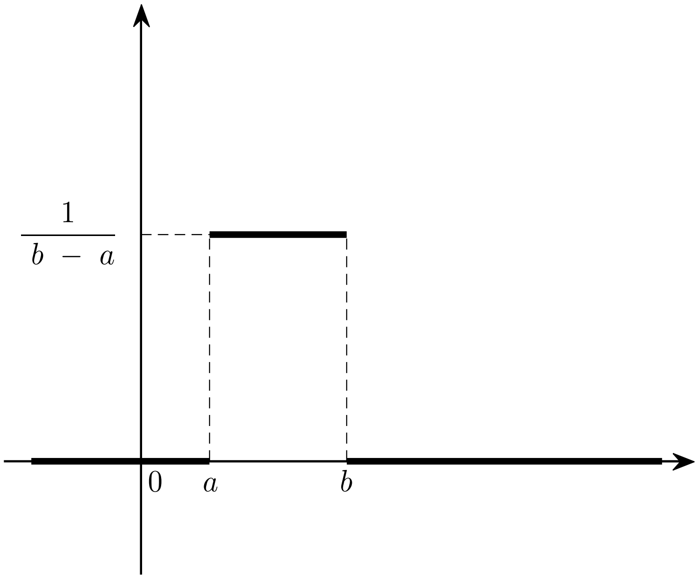


Fig. 8: Uniform Distribution (multimodal)

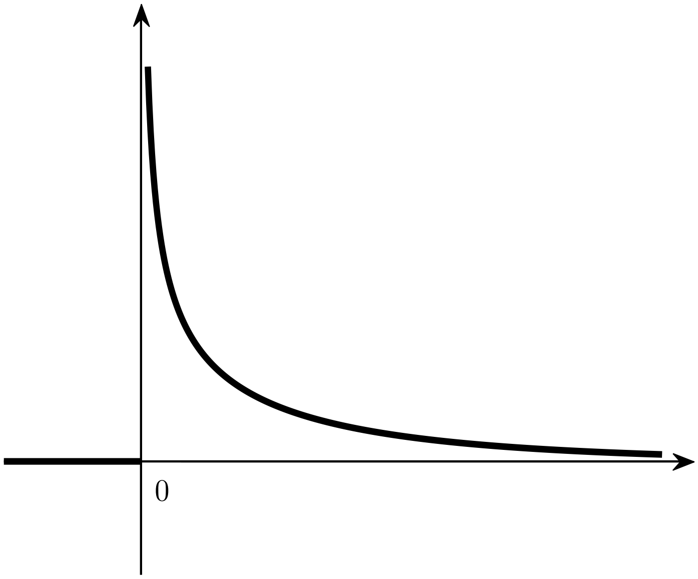


Fig. 9: χ^2 Distribution (no mode)

In general,

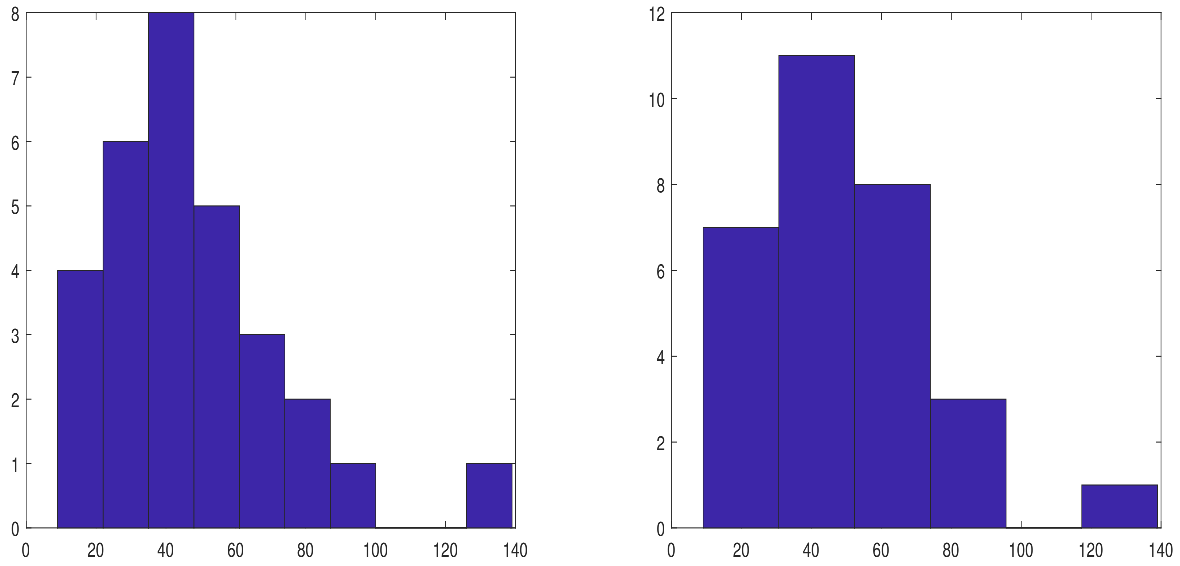
$$x_{mo} \approx \overline{x} - 3(\overline{x} - \overline{M}).$$

This empirical formula was given by K. Pearson.

Example 4.15. 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 10(a)). If we have only 6 classes, then the second one is the modal class, $[30.7, 52.4)$, with mark 41.55 (Figure 10(b)).



(a) $n = 10$ bins

(b) $n = 6$ bins

Fig. 10: Modal class