

Chapter 1

Descriptive Statistics

1.1 Descriptive Statistics of One-Dimensional Data

Descriptive statistics to use with the representation of data sets (compilations of various data). It uses certain numbers to characterize such data sets (for example the sample mean) and represents them graphically in a coordinate system. The first deal with *one-dimensional* data, where *one* measurement quantity of a test object is determined. We will use the following example to study important concepts and procedures more precisely.

1.1.1 Measuring the Melting Heat of Ice

As an introductory example, let us consider two *data sets* with which to compare two methods of determining the latent heat of melting ice.

Example 1.1.1

Repeated measurements of heat released when passing from ice at -0.7°C to water at 0°C yielded the values (in cal/g) listed in Table 1.1. Even though the measurements were taken with the greatest possible care and all interference signals were disabled, the measurement values vary with either method. The following questions then arise:

- Is there a difference between method *A* and method *B*?
- If yes, how can we determine this difference?

It is evident that both methods exhibit measurement values around 80. However, method *A* has only 2 values out of 13 *below* 80, while method *B* has only 2 values out of 8 *above* 80. So, the values from method *A* are more likely to be larger than those from method *B*. But what does “more likely” mean here? So, it is worth summarizing the two measurement series in such a way as to compare the two methods. ◀

Chapter 1 Descriptive Statistics

Method A	79.98	80.04	80.02	80.04	80.03	80.03	80.04	79.97	80.05
Method A	80.03	80.02	80.00	80.02					
Method B	80.02	79.94	79.98	79.97	79.97	80.03	79.95	79.97	

Table 1.1: Measurements to determine the latent heat of melting ice based on two methods.

Descriptive statistics is about how (quantitative) data can be organized and summarized. The objective is simplifying the interpretation and subsequent statistical analysis of the data. We do this using:

- graphic representations
- summaries of data, which should highlight the important characteristics of the data, such as the mean location of the measurement values and their dispersion around the mean location.

These so-called *summary statistics* should roughly characterize the data and summarize it numerically.

When performing statistical analyses, as we will study them throughout this course, it is extremely important to avoid simply adapting a model or applying a statistical method blindly. The data should always be represented using appropriate graphical tools *and* the summary statistics, since this is the only way to find (possibly unexpected) structures and features.

The data below is designated x_1, \dots, x_n , where n is called the *size* of the measurement series. In the case of the measurement series for method A, we have $n = 13$:

$$x_1 = 79.98, \quad x_2 = 80.04, \quad \dots, \quad x_{13} = 80.02$$

1.1.2 Representing Measurement Values

Before we discuss summary statistics and graphic representations of data sets, we must settle some rules for representing measurement values. To this end, we need the concepts of *decimal places* and *significant digits*.

Decimal places refer to the digits used in the decimal representation of a number that are to the right of the decimal point. In the example above, the measurement points

$$x_1 = 79.98, \quad x_2 = 80.04, \quad \dots, \quad x_{13} = 80.02$$

have *two* decimal places.

Significant digits are defined as the first non-zero digits through the rounded digit. The rounded digit is the last digit that can be indicated after rounding. So, in the example above, we have *four* significant digits.

Example 1.1.2

<i>Number</i>	<i>Number of significant digits</i>	<i>Number of decimal places</i>
98.76	4	2
0.009876	4	6
$987.6 \cdot 10^4$	4	1
$9.876 \cdot 10^6$	4	3



Remarks:

- i. Integers have no decimal places.
- ii. In some cases, determining the significant digits can be ambiguous: Does 20 have one, two, or even more significant digits? Depending on the context, a number should be considered exact if it is used as a natural number, for example, or it should be considered rounded if it is used as a numerical value of a physical quantity. For an exact number, the question of significance does not come up, since it can be extended with any number of zeros after the decimal point.
- iii. Using the scientific notation with powers of 10 can prevent ambiguity of a quantity received from measurement technology with the numerical value 20. So, in the case of one significant digit, this would be $2 \cdot 10^1$; in the case of three significant digits, $2.00 \cdot 10^1$.



Representing Computational Results

The two following rules apply to the representation of a computational result of measurement values:

1. The result of an *addition/subtraction* has the same number of decimal places as the number with the fewest decimal places.
2. The result of a *multiplication/division* has the same number of significant digits as the number with the fewest significant digits.

Example 1.1.3

<i>Number</i>	<i>Lowest number of significant digits</i>	<i>Lowest number of decimal places</i>	<i>Result</i>
$20.567 + 0.0007$		3	20.568
$12 + 1.234$		0	13
$12.00 + 1.234$		2	13.23
$12.000 + 1.234$		3	13.234
$1.234 \cdot 3.33$	3		4.11
$1.234 \cdot 0.0015$	2		0.0019



Remark 1.1.1

Rounding should occur as late as possible during the calculation. Otherwise, multiple rounding errors could combine into a very large total deviation. To prevent this amplification, known quantities should be populated with at least one more digit in the intermediate calculations than are shown in the final result.

1.1.3 Summary Statistics

It often makes sense to summarize data sets *numerically*. This reduces data sets to one or more numbers. We usually use two *parameters*: One describes the mean location of the measurement values and the other the variability or dispersion of the same measurement values. By dispersion, we mean the measurement values' "average" deviation from the mean location.



Arithmetic Mean

The best known quantity for the mean location is the widely known average, or

Arithmetic Mean \bar{x}

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n} \sum_{i=1}^n x_i$$

Remark 1.1.2

In some cases, the notation \bar{x}_n may be used where n denotes the size of the measurement series.

Example 1.1.4 Measuring the melting heat of ice with method A

The arithmetic mean of the $n = 13$ measurements is

$$\bar{x}_{13} = \frac{79.98 + 80.04 + \cdots + 80.03 + 80.02 + 80.00 + 80.02}{13} = 80.02$$

So, we add up all the values and divide the sum by the number of values.

With **Python** we calculate the sample mean as follows:

```
from pandas import Series, DataFrame
import pandas as pd

methodA = Series([79.98, 80.04, 80.02, 80.04, 80.03, 80.03,
80.04, 79.97, 80.05, 80.03, 80.02, 80.00, 80.02])
print(methodA.mean())

methodA.mean()
```



Empirical Variance and Standard Deviation

Even though the arithmetic mean already says a lot about the data set, it only describes partially describes it. For example, let us consider the following two data sets of (fictitious) student grades:

2; 6; 3; 5 und 4; 4; 4; 4

Chapter 1 Descriptive Statistics

Both have the same sample mean of 4, but they have very different distributions of the data around that mean. In the first case, there are two good and two poor students; in the second case, all students are equally good. We say the data sets have different dispersions around the means.

We want to record this dispersion numerically. A first approach consists in taking the average of the *distances from the mean*. In the first case, this would be

$$\frac{(2 - 4) + (6 - 4) + (3 - 4) + (5 - 4)}{4} = \frac{-2 + 2 - 1 + 1}{4} = 0$$

In the second case, we also obtain 0. So, this method does not contribute much to describing the dispersion, since the distances from the mean can be *negative* and can cancel each other out, as in the case above.

The next approach consists in replacing the distances from the mean with the absolute values of the distances from the mean. In the first case, we then obtain

$$\frac{|(2 - 4)| + |(6 - 4)| + |(3 - 4)| + |(5 - 4)|}{4} = \frac{2 + 2 + 1 + 1}{4} = 1.5$$

The grades deviate on average by 1.5 points from the sample mean. In the second case, the value is obviously 0. The greater this value becomes (and it is always greater than or equal to 0), the more the data points diverge while having the same mean. This value for the dispersion is also known as the *mean absolute deviation*.

Since absolute value are not easy to work with (e.g. in derivatives), the (seemingly more complex) *empirical variance* and *empirical standard deviation* for the measure of variability or dispersion of the measurement values. These are defined as follows:

Empirical Variance $\text{Var}(x)$ and Standard Deviation s_x

$$\text{Var}(x) = \frac{(x_1 - \bar{x}_n)^2 + (x_2 - \bar{x}_n)^2 + \cdots + (x_n - \bar{x}_n)^2}{n - 1} = \frac{1}{n - 1} \sum_{i=1}^n (x_i - \bar{x}_n)^2$$

$$s_x = \sqrt{\text{Var}(x)} = \sqrt{\frac{1}{n - 1} \sum_{i=1}^n (x_i - \bar{x}_n)^2}.$$

Remarks:

- For the variance, we square the deviations $x_i - \bar{x}_n$ so the deviations from the mean do not cancel each other out. The denominator $n - 1$, instead of n , is mathematically justified.

- ii. The standard deviation is the square root of the variance. Since our calculation of the variance uses the squares of the distances from the mean, the square root brings us back to the same unit as in the data. The value of the empirical variance has no meaning in the physical world. We simply know the greater this value, the greater the dispersion. ♦

Example 1.1.5 Measuring the melting heat of ice with method A

The arithmetic mean of the $n = 13$ measurements is $\bar{x}_{13} = 80.02$ (see above), and the empirical variance gives us

$$\begin{aligned}\text{Var}(x) &= \frac{(79.98 - 80.02)^2 + (80.04 - 80.02)^2 + \cdots + (80.00 - 80.02)^2 + (80.02 - 80.02)^2}{13 - 1} \\ &= 0.0005744\end{aligned}$$

Then the empirical standard deviation is

$$s_x = \sqrt{0.000574} = 0.02397$$

Therefore, “the average” deviation from the mean is 0.023 97 cal/g.

For method B, we find $\bar{x}_8 = 79.98$ and $s_x = 0.03137$ with an analogous interpretation.



The empirical variance and standard deviation are arduous to compute, which is why we use **Python**. For the variance, we use the method

```
methodeA.var()
```

and for the standard deviation, we use

```
methodeA.std()
```

1.1.4 Other Summary Statistics

Below, we study two alternative parameters, namely the *median* as the measure of location and the *interquartile range* as the measure of dispersion.

Median

Another measure of the average location is the *median*, which is the value at which about half the measurement values fall below it. For example, if the median for a test is 4.6, then half the class received a grade below 4.6. Conversely, the other half has grades *above* the median.

To determine the *median*, we must first order the data by size:

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$$

For the data in method *A*, this yields

79.97; 79.98; 80.00; 80.02; 80.02; 80.02; 80.03; 80.03; 80.03; 80.04; 80.04; 80.04; 80.05

In **Python**, you sort the data by size as follows:

```
methodeA.sort_values
```

Then, the median of these 13 measurements is the value of the middle observation. In this case, it is the value of the 7th observation:

79.97; 79.98; 80.00; 80.02; 80.02; 80.02; **80.03**; 80.03; 80.03; 80.04; 80.04; 80.04; 80.05

The median of the data set from method *A* is therefore 80.03. 6 observations are less than or equal to 80.03, and 6 measurement values are greater than or equal to 80.03. In this example, the number of data points is odd. There is therefore the middle observation. If the number of data points is even, then there are effectively two middle observations. In this case, we use the mean of the two middle observations is the median. The data sets for method *B* has 8 observations. We order the data set by size and defined as the median the average of 4th and 5th observations:

79.94; 79.95; 79.97; **79.97; 79.97**; 79.94; 80.02; 80.03

$$\frac{79.97 + 79.97}{2} = 79.97$$

The median of the data from method *B* is 79.97. This means half the measurement values are less than or equal to this value, and the other half is greater than or equal to it. The two middle observations here happen to have the same value, but this is generally not the case.

Example 1.1.6

Using **Python**, we determine the median as follows for method *A*

```
methodeA.median()
```

and for method *B*


```
methodeB = Series([80.02, 79.94, 79.98, 79.97, 79.97, 80.03,
79.95, 79.97])

methodeB.median()
```



We now have two measures of location for the middle of a data set: the arithmetic mean and the median. What are the advantages of each measure? One property of the median is its *robustness*. Extreme observations have less of an impact on the median than on the arithmetic mean.

Example 1.1.7 Measuring the melting heat of ice with method A

A typo occurred with the largest observation ($x_9 = 80.05$), and $x_9 = 800.5$ was entered. Then the arithmetic mean becomes

$$\bar{x}_{13} = 135.44$$

But the median remains the same as earlier:

$$x_{(7)} = 80.03$$

So, changing an observation greatly impacts the arithmetic mean, while the median remains the same here – the latter is *robust*.



Quartiles

The *lower quartile* is the value at which 25 % of all observations are less than or equal to this value, and 75 % are greater than or equal to this value. Respectively, the *upper quartile* is the value at which 75 % of all observations are less than or equal to this value, and 25 % are greater than or equal to this value.



However, most data sets do not have *exactly* 25 % of the number of observations. Unfortunately, there are several different ways to handle these cases.

Example 1.1.8

Method A has $n = 13$ measurement points; 25 % of this number is 3.25.

- We *choose* the next greater value $x_{(4)}$ as the lower quartile. Then, a little bit more than 25 % of the values are less than or equal to this value:

79.97; 79.98; 80.00; **80.02**; 80.02; 80.02; 80.03; 80.03; 80.03; 80.04; 80.04; 80.04; 80.05

The lower quartile is therefore 80.02. About one quarter of the measurement values are less than or equal to this value.

- However, we *could* also choose the next smaller value, since 3.25 is closer to 3 than to 4, and then the lower quartile was 80.00.
- We *could* also take the mean value of the third and the fourth value, the yielded 80.01.
- We *could* also interpolate linearly. Then we take 0.25 of the difference of the third and fourth value and obtain as lower quartile 80.005.



There are also other variants of definitions of the lower and upper quartile. Depending on the software, different versions are implemented.

Quartiles

The *lower quartile* is the value, for which *around* 25 % of all observations are less than or equal to it, and *around* 75 % are greater than or equal.

Accordingly, the *upper quartile* is the value, for which *around* 75 % of all observations are less than or equal to it, and 25 % are greater than or equal.

Here, the word “around” means “as close to as possible”.

The software **Python** does not have specific commands for quartiles. We can however use the more general **quantile** command (we will meet quantiles a little bit later).

Example 1.1.9

For the lower quartile of method A, the command reads

```
methodeA.quantile(q=.25)
```

and so 25 % of the values are less than or equal to 80.02, and 75 % are greater than or equal to 80.04.

For the upper quartile, we have

```
methodeA.quantile(q=.75)
```

So 75 % of the values are less than or equal to 80.04, and 25 % are greater than or equal to 80.04.

For method B we obtain

```
methodeB.quantile(q=.25)
```



Remarks:

- i. The different definitions yield different values for the quartiles as we saw in example 1.1.8.
- ii. However, most of the times the values for the quartiles according to the different definitions are very close to each other, and their interpretation is roughly the same.
- iii. Due to robustness of the quartiles, using a different definition moves them by at most one value of the data series. For large data sets, this is usually irrelevant.



Example 1.1.10

We consider the different definitions for the quartiles, which are implemented in **Python**. They are called with the option `interpolation=...`. For the details of the several definitions, we refer to the help function in Ipython.

For the lower quartile of method *A*, we obtain

```
methodeA.quantile(q=.25, interpolation="linear")
methodeA.quantile(q=.25, interpolation="lower")
methodeA.quantile(q=.25, interpolation="higher")
methodeA.quantile(q=.25, interpolation="midpoint")
methodeA.quantile(q=.25, interpolation="nearest")
```

In this case, the different definitions yield no differences at all.

For the lower quartile of method *B*, we obtain

```
methodeB.quantile(q=.25, interpolation="linear")
methodeB.quantile(q=.25, interpolation="lower")
methodeB.quantile(q=.25, interpolation="higher")
methodeB.quantile(q=.25, interpolation="midpoint")
methodeB.quantile(q=.25, interpolation="nearest")
```

Here, we have very small differences, and the interpretation do not change considerably. About 25% of the values are less than or equal to 79.97 (or 79.95 or 79.96, respectively). The values are very close to each other. ◀

Interquartile Range

The interquartile range

$$\text{upper quartile} - \text{lower quartile}$$

is another measure of the data's dispersion. It measures the length of the interval that contains about the middle half of the observations. The smaller this measure, the closer this half of all values is to the median and the smaller the dispersion. This measure of dispersion is robust.

Example 1.1.11

The interquartile range with method *A* is

$$80.04 - 80.02 = 0.02$$

```
q75, q25 = methodeA.quantile(q = [.75, .25])  
iqr = q75 - q25  
iqr
```

So, about half of all measurement values are within a range of length 0.02. ◀

Quantiles

The concept of *quantiles* allows us to generalize the concept of quartiles to any other percentage. The 10 % quantile is thus the value at which 10 percent of the values are less than or equal to this value and 90 % of the values are greater than or equal to it.

Quantile

The *empirical* α -quantile corresponds to the value at which $\alpha \times 100\%$ of the data points are less than or equal to this value and $(1 - \alpha) \times 100\%$ of the points are greater than or equal to it.

Remarks:

- i. The empirical median is the empirical 50 %-quantile; the empirical 25 %-quantile is the lower quartile; the empirical 75 %-quantile is the upper quartile.
- ii. The exact definition of the empirical α -quantile is:

$$\frac{1}{2}(x_{(\alpha n)} + x_{(\alpha n + 1)}) \text{ , if } \alpha \cdot n \text{ is an integer,}$$

$$x_{(k)} \text{ where } k \text{ is the number } \alpha \cdot n \text{ rounded up , if } \alpha \cdot n \notin \mathbb{N}.$$



Example 1.1.12 Measuring the melting heat of ice with method A

We determine the median and the lower and upper quartiles based on the definition above.

There are $n = 13$ measurement values, which we first have to order by size: the smallest value is $x_{(1)} = 79.97$; the third largest value $x_{(3)} = 80.00$; the largest value $x_{(13)} = 80.05$. We want to determine the 25 %-quantile, the median, and the 75 %-quantile. In the case of 25 %-quantile, we then have $\alpha = 0.25$, so

$$\alpha \cdot n = 0.25 \cdot 13 = 3.25$$

which is not an integer; hence we round 3.25 up to 4 and obtain for the 25 %-quantile $x_{(4)} = 80.02$. In the case of the median, we have $\alpha = 0.5$, so

$$\alpha \cdot n = 0.5 \cdot 13 = 6.5$$

which is not an integer; hence we round 6.5 up to 7 and obtain for the the median $x_{(7)} = 80.03$. In the case of 75 %-quantile, we have $\alpha = 0.75$, so

$$\alpha \cdot n = 0.75 \cdot 13 = 9.75$$

which is not an integer; hence we round 9.75 up to 10 and obtain for the 75 %-quantile the observed value $x_{(10)} = 80.04$. ◀

Example 1.1.13

We calculate the 10 %- and 70 %-quantile from method A as follows:

```
methodeA.quantile(q=.1)
methodeA.quantile(q=.7)
```

About 10 % of the measurement values are less than or equal to 79.98. Accordingly, about 70 % of the measurement values are less than or equal to 80.03. ◀

Example 1.1.14

A class of 24 students received the following grades on a test:

4.2, 2.3, 5.6, 4.5, 4.8, 3.9, 5.9, 2.4, 5.9, 6, 4, 3.7, 5, 5.2, 4.5, 3.6, 5, 6, 2.8, 3.3, 5.5, 4.2, 4.9, 5.1

We now use **Python** to calculate various quantiles:

```
noten = Series([4.2, 2.3, 5.6, 4.5, 4.8, 3.9, 5.9, 2.4, 5.9,
                6, 4, 3.7, 5, 5.2, 4.5, 3.6, 5, 6, 2.8, 3.3,
                5.5, 4.2, 4.9, 5.1])

noten.quantile(q = np.linspace(start=.2, stop=1, num=5))
```

So, about 20 % of the students received a 3.6 or worse. Exactly 20 % of the students is not possible, since would mean 4.8 students. The 60 %-quantile means 60 percent of the students received a 4.9 or worse. Hence 40 % received a 4.9 or better. ◀

Remark 1.1.3

Also here, the statements do not change much, when we say that about 20 % of the students had 3.7 or worse.

1.1.5 Graphical Methods

Histogram



A so-called *histogram* gives us a graphical overview of the values that occur. Histograms help answer the question of what value range has a particularly high number of data points. If there are many data points, then it does not make sense to observe each value individually. We form so-called *bins*, each of which represents a section of the observation area. To draw a histogram, we build classes (for the sake of convenience, of equal width) and count the number of observations fall in each class. There are different types of histograms; here, we only work with the most common type.

Example 1.1.15

In Figure 1.1, we see a histogram of the results of an IQ test involving 200 people.

- The classes' width was set to 10 IQ points; it is the same for every class.
- The bars' height indicates the number of people whose score falls within the respective class. For example, approx. 14 people are in the class of IQ points between 120 and 130.

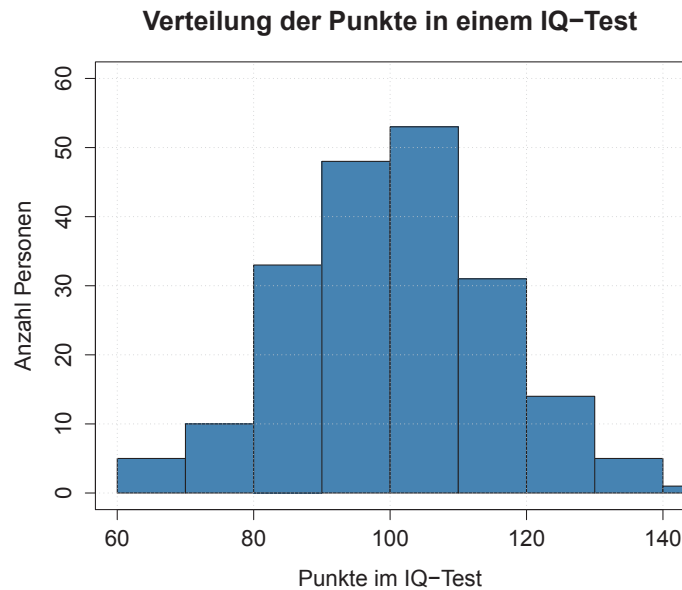


Figure 1.1: Histogram of the IQ-test results of 200 persons.



Ste-by-step construction of a histogram:

- Divide the volume of data into classes. There are a few rules of thumb available to set the number of classes or rectangles: For fewer than 50 measurements, use 5 to 7 classes; for more than 250 measurements, choose 10 to 20 classes¹. In the simplest case, all classes have the same width, although this is not necessary.
- We then draw a bar for each class, whose height is proportional to the number of observations in that class.
- We divide the number of observations in a class by the total number of observations, which gives us the percentage of a class out of the total observation.

Example 1.1.16

For method A, the histogram of is given in Fifure 1.2. It was generated using the following code:

```
import pandas as pd
from pandas import DataFrame, Series
import matplotlib.pyplot as plt
```

¹If necessary, we can also calculate the number of classes k based on Sturges' formula: $k = 1 + \log_2 n = 1 + 3.3 \cdot \log_{10} n$, where n is the number of measurements.

Chapter 1 Descriptive Statistics

```
methodA = Series([79.98, 80.04, 80.02, 80.04, 80.03, 80.03,
80.04, 79.97, 80.05, 80.03, 80.02, 80.00, 80.02])

methodB = Series([80.02, 79.94, 79.98, 79.97, 79.97, 80.03,
79.95, 79.97])

methodA.plot(kind="hist", edgecolor="black")

plt.title("Histogram of method A")
plt.xlabel("methodA")
plt.ylabel("Frequency")

plt.show()
```

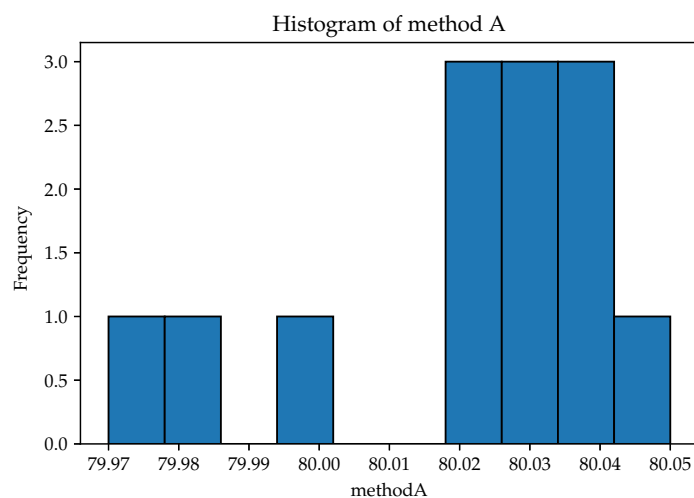



Figure 1.2: Histogram of the melting heat measured with method A

Remarks:

- pandas** itself has no graphical capabilities, but falls back on the library **matplotlib**.
- We used the general **pandas** method **plot** and then choose the option **... (kind = "hist", ...)** for histogram.
- pandas** chooses by default 10 intervals (bins). That can, for example, be changed to 7 bins **... (kind="hist", ..., bins=7)**.
- Meaning of the numbers (frequency): Because 10 bins were chosen and the data values lie in the interval $[79.97, 80.05]$, then the bin width is $(80.05 - 79.97) / 10 = 0.008$.

The 1st class, from 79.97 to 79.978, accounts for the observations with values 79.97; the 2nd class, 79.98; etc.

- v. What happens, if a value is on the edge of a bin? The value could be accounted to the left or to the right bin (but not to both). Depending on the choice, the histogram would look a little different. Such considerations have little incidence on large data sets.
- vi. The **pandas** command also allows from setting the number and width of classes, changing the labels, etc. (see the exercises). 



Example 1.1.17

In the histogram above, the height of the bars corresponds directly to the observations in a class. It is often better and clearer to choose the height of the bars such that the surface of each bar corresponds to the respective observations' percentage out of the total number of observations. The total area of the bars must then be equal to 1. Then we speak dann of a *normed histogram*, in which the height of the bars represent the *density* of the values. This is shown in Figure 1.3 for method A. It was plotted using the option **density=True**:



```
methodeA.plot(kind="hist", density=True, edgecolor="black")

plt.title("Normed histogram of method A")
plt.xlabel("methodA")
plt.ylabel("Density")

plt.show()
```

The vertical axis now shows the densities. We can thus read that more than $(80.018 - 80.026) \cdot 28.846 = 0.23$, so about 23 % of the data is between 80.018 and 80.026.

The bar height is determined by multiplying the number of observations included in the bar by $\frac{1}{n}$, where n is the total number of observations, and then dividing this result by the bar width. In our example, 3 observations lie in the interval 80.018 und 80.026, then height of this bar is

$$\frac{\frac{1}{13} \cdot 3}{0.008} = 28.8462$$



The normed histogram has the advantage that measurements with different volumes are easier to compare. So, if we were to now measure 30 observations with method A, the distributions of measurement values into their respective classes would be easier to compare with the densities.

Chapter 1 Descriptive Statistics

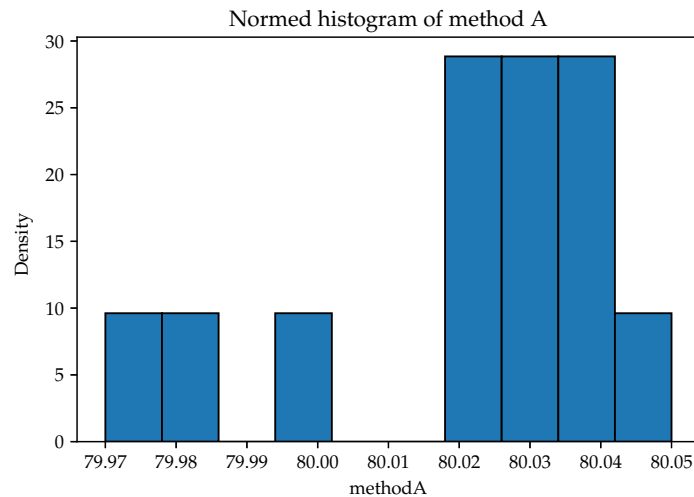


Figure 1.3: Normed histogram of method A

Box Plot

A *box plot* (see Figure 1.4) consists of

- a rectangle whose height is bound by the empirical 25 %- and 75 %-quantile,
- lines leading from this rectangle to the smallest and largest “normal” values (based on the definition, a “normal” value is at most 1.5 times the interquartile range away from the two aforementioned quartiles),
- a horizontal line for the median,
- small circles marking the outliers.

Example 1.1.18

The boxplot of method A looks as follows:

```
methodeA.plot(kind="box", title="Boxplot Methode A")
```

Remarks:

- i. Half of the observations are between the upper quartile, 80.04, and the lower quartile, 80.02, with an interquartile range of 0.02.
- ii. The median is at 80.03.
- iii. The “normal” range of values is between 80.00 and 80.05.

Chapter 1 Descriptive Statistics

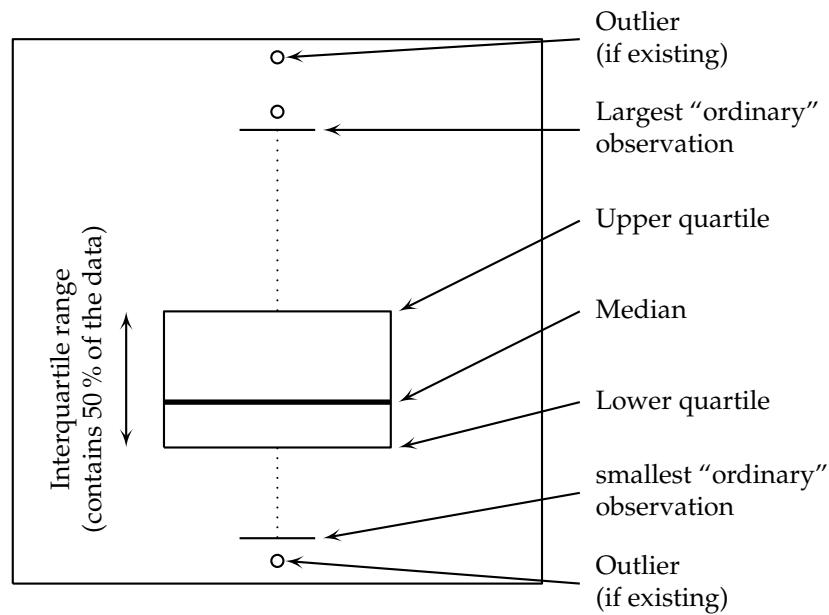


Figure 1.4: Box Plot

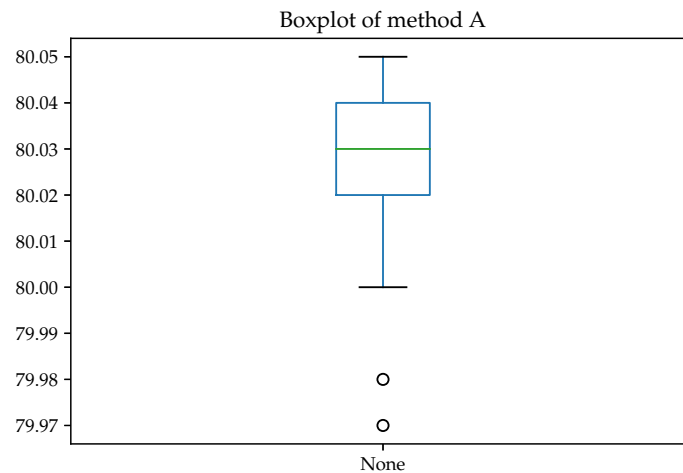


Figure 1.5: Boxplot of method A

- iv. We have two outliers, 79.97 and 79.98.
- v. We had already calculated points i) and ii) when dealing with the quantiles. The box plot thus represents our calculations graphically. ♦

So, the box plot is primarily suitable when comparing the distributions of the data in different groups (which generally correspond to different test conditions).

Example 1.1.19

In our introductory example, we used two methods to determine the melting heat. We can also contrast the box plots and compare the methods with one another, see Figure 1.6. It was produced by creating from the two **Series** a **DataFrame** with the name **methode** and calling its method **plot**.

```
methode = DataFrame({
    "methodeA": methodeA,
    "methodeB": methodeB
})

methode.plot(kind="box", title="Boxplot von Methode A und B")
```

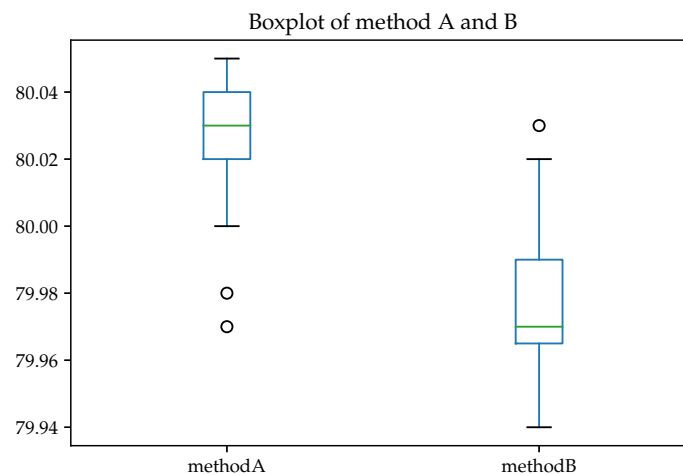



Figure 1.6: Boxplots for the two methods to determine the melting heat of ice.

Remarks:

- i. Method *A* produces larger values than method *B*, since the median from *A* is larger.
- ii. The data from method *A* has less dispersion than the data from method *B*, since the rectangle is smaller (interquartile range!).

Empirical Cumulative Distribution Function

Another graphical representation of the data comes through the *empirical cumulative distribution function*. It has the advantage over a histogram of making the median very easy to read. 

The *empirical cumulative distribution function* $F_n(\cdot)$ is a step function generated as follows: The function is equal to zero to the left of $x_{(1)}$, and at every $x_{(i)}$ a jump occurs whose height is $1/n$ (if a value occurs several times, the jump's height is the corresponding multiple of $1/n$). This procedure is executed tangibly in the following example.

Example 1.1.20 Method A for melting heat

Figure 1.7 plots the empirical cumulative distribution function of method A. It is constructed as follows:

- To the left of 79.97, the function is 0, since there are no smaller observed values.
- At 79.97, the function jumps to $n = 1/13 \approx 0.077$.
- The function then stays constant until 80.00, since there are no other observed values before then. At 80.00, the function jumps up again by 0.077, because there is a measurement value there.
- At 80.02, the function jumps up again by $3 \cdot 0.077$, because there are 3 observed values there etc.
- At 80.05, the function jumps one last time, and the function takes the value 1.



What can we discern from the cumulative distribution function?

Example 1.1.21

- At 0.5 along the vertical axis, we have added up exactly half of all values, see Figure 1.8. If we draw a horizontal line from 0.5 (see the green line in Figure 1.8), it crosses the cumulative distribution function at 80.03, which matches the median exactly.
- Where the cumulative function makes a high jump, it has many observed values. That means most observed values here are between 80.02 and 80.04. But the values match the lower and upper quartiles exactly. The function is compared to the associated box plot.



Chapter 1 Descriptive Statistics

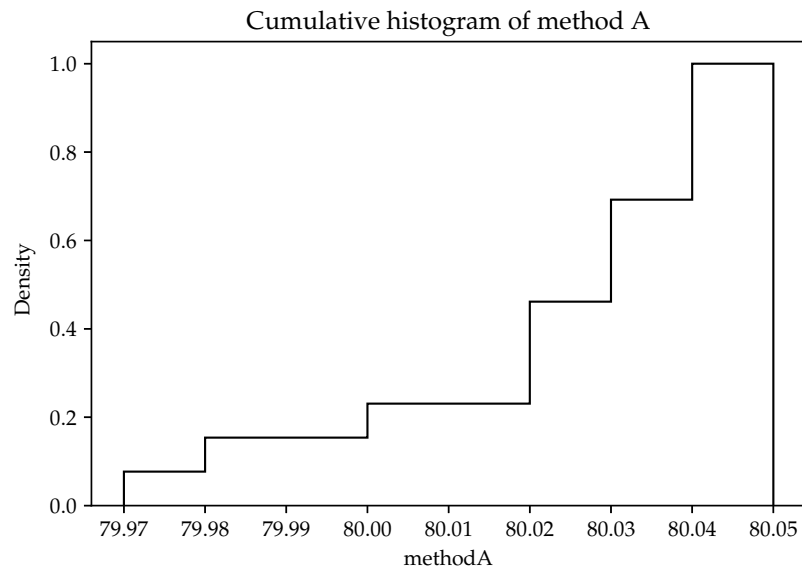


Figure 1.7: Empirical cumulative distribution function of the measurements for the melting heat of ice using method A.

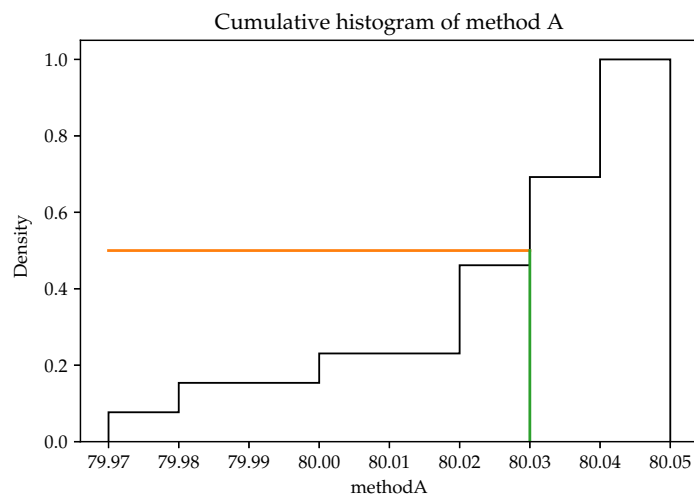


Figure 1.8: Empirical cumulative distribution function of the measurement of the melting heat of method A.

General

The **empirical cumulative distribution function** is defined as

$$F_n(a) = \frac{1}{n} \text{Anzahl}\{i \mid x_i \leq a\}.$$

Example 1.1.22

With **Python**, the cumulative distribution function in Figure 1.8 can be plotted in the following manner:

```
methodeA.plot(kind="hist", cumulative=True, histtype="step",  
density=True, bins=8, edgecolor="black")
```

Strictly speaking, this is a cumulative histogram, so it depends on the number and width of the bins. It does not correspond exactly our definition of a cumulative distribution function, but we do not go into these details here. ◀

1.2 Descriptive Statistics of Two-Dimensional Data

With two-dimensional data, *two* different quantities are measured on each test object. For example, the height *and* weight is measured on each person within a group.

Example 1.2.1 Wine Consumption and Mortality

As an example, let us consider a data set (see Table 1.2) that includes the average wine consumption (in liters per person per year) and the death rate (mortality) from heart and cardiovascular diseases (number of deaths per 1000 people aged 55 to 64) in 18 industrialized countries ². The question arises of whether this data suggests there is a relationship between the death rate from cardiovascular disease and wine consumption. A quick look at the table shows a higher wine consumption seems to lead to fewer deaths due to heart and cardiovascular diseases. ◀

1.2.1 Graphical Representation: Scatter Plot

Graphical representation is an important step in the study of two-dimensional data, usually using a so-called *scatter plot*. This involves interpreting and representing two measurements in each case as coordinates of points in a coordinate system.

²A. S. St. Leger, A. L. Cochrane, and F. Moore, “Factors Associated with Cardiac Mortality in Developed Countries with Particular Reference to the Consumption of Wine.” *Lancet*, 1979

Country	Wine consumption	Mortality from heart disease
Norway	2.8	6.2
Scotland	3.2	9.0
Great Britain	3.2	7.1
Ireland	3.4	6.8
Finland	4.3	10.2
Canada	4.9	7.8
United States	5.1	9.3
Netherlands	5.2	5.9
New Zealand	5.9	8.9
Denmark	5.9	5.5
Sweden	6.6	7.1
Australia	8.3	9.1
Belgium	12.6	5.1
Germany	15.1	4.7
Austria	25.1	4.7
Switzerland	33.1	3.1
Italy	75.9	3.2
France	75.9	2.1

Table 1.2: Wine consumption (liters per person per year) and mortality from cardiovascular disease (deaths per 1000) in 18 countries.

Example 1.2.2

In our example, a country represents an experimental unit, and the parameters “wine consumption” x_1, \dots, x_{18} and “mortality” y_1, \dots, y_{18} are measured. If we write the data in the format $(x_1, y_1), \dots, (x_{18}, y_{18})$, we are primarily interested in the relationships and dependencies between the variables x and y . The *scatter plot* shows the dependencies between the two measurement quantities, representing the data as points on the plane: The i -th observation (i -th country) corresponds to the point with coordinates (x_i, y_i) . Figure 1.9 displays the scatter plot for the measurement quantities “wine consumption” $(x_1, x_2, \dots, x_{18})$ and “mortality” $(y_1, y_2, \dots, y_{18})$. We see a monotonously decreasing relationship: Countries with high wine consumption tend to have a lower mortality rates from heart and cardiovascular diseases.

The scatterplot in Figure 1.9 was created in **Python** as follows.

```
import pandas as pd
from pandas import DataFrame, Series
import numpy as np

mort = DataFrame({
    "wine": ([2.8, 3.2, 3.2, 3.4, 4.3, 4.9, 5.1, 5.2, 5.9, 5.9,
```


Chapter 1 Descriptive Statistics

```
        6.6, 8.3, 12.6, 15.1, 25.1, 33.1, 75.9, 75.9]),  
    "mor": ([6.2, 9.0, 7.1, 6.8, 10.2, 7.8, 9.3, 5.9, 8.9, 5.5,  
            7.1, 9.1, 5.1, 4.7, 4.7, 3.1, 3.2, 2.1])  
))  
  
mort.plot(kind="scatter", x="wine", y="mor")  
  
plt.xlabel("Wine consumption (liter per person per year)")  
plt.ylabel("Mortality")  
  
plt.show()
```

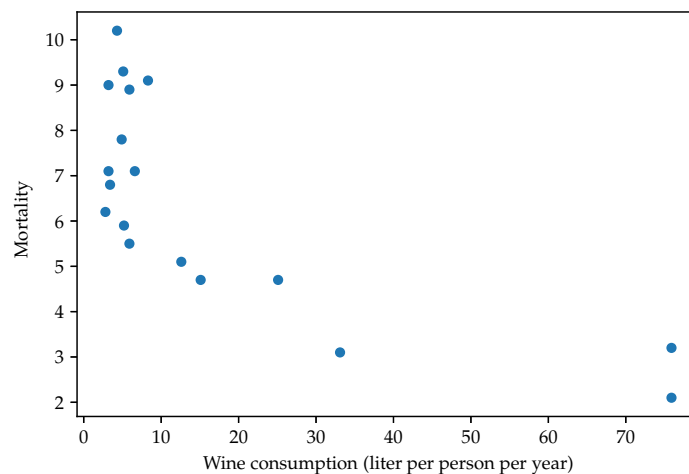


Figure 1.9: Scatter plot of the mortality and the wine consumption in 18 industrialized countries.

Remarks:

- i. The conclusion that elevated one consumption is health is rash and supposedly *false*. It appears that a higher wine consumption leads to fewer deaths due to heart and cardiovascular diseases. The influence of the higher wine consumption on other organs (e.g. the liver) or on the number of traffic accidents is *not* studied here.
- ii. Although a relationship between wine consumption and mortality can be *surmised* from the scatterplot, a *causal* relationship does not necessarily exist between the two quantities. ♦

1.2.2 Simple Linear Regression

In the example above, we were able to determine a negative dependence (the more... the less...) between mortality and wine consumption. Such a dependency can often be assumed to be very simple, namely *linear*.

Example 1.2.3 Relationship between the number of pages and the price of a Book



Let us now explain the simple linear regression model with a fictitious example. The thicker the novel (hardcover), the more expensive it generally is. So, there is a relationship between the number of pages x and the book's price y . We go into a bookstore and take 10 novels of different thicknesses. We thus take one book each with 50, 100, 150, ..., 450, 500 pages. We make a note of each book's number of pages and price. Using this data, we create Table 1.3.

The table does make it evident that the books tend to cost more. If we had a formula for the relationship between the book price and the number of pages, we would be able to predict the price of books we had not observed based on their page numbers. So, how much would a book with 375 pages likely cost? Or we could find out how much a book with "zero" pages would cost. This would refer to the publisher's basic costs incurred regardless of the number of pages: binding, administrative expenses for every book, etc. How could we describe this relationship with the formula? The scatter plot in Figure 1.10 distinctly illustrates this relationship graphically.

```
book = DataFrame({
    "pages" : np.linspace(50,500,10),
    "price" : [6.4, 9.5, 15.6, 15.1, 17.8, 23.4,
              23.4, 22.5, 26.1, 29.1]
})

book.plot(kind="scatter", x="pages", y="price")

plt.xlabel("Pages")
plt.ylabel("Price")

plt.show()
```

At first glance, a straight line seems to fit the data quite well. Such a line would have the format

$$y = a + bx$$

where y is the book's price and x its number of pages. Parameter a then describes the publisher's basic costs, and parameter b refers to the cost per page. ◀

Chapter 1 Descriptive Statistics

	Number of pages	Book price (CHF)
Book 1	50	6.4
Book 2	100	9.5
Book 3	150	15.6
Book 4	200	15.1
Book 5	250	17.8
Book 6	300	23.4
Book 7	350	23.4
Book 8	400	22.5
Book 9	450	26.1
Book 10	500	29.1

Table 1.3: Relationship between book price and number of pages (fictitious).

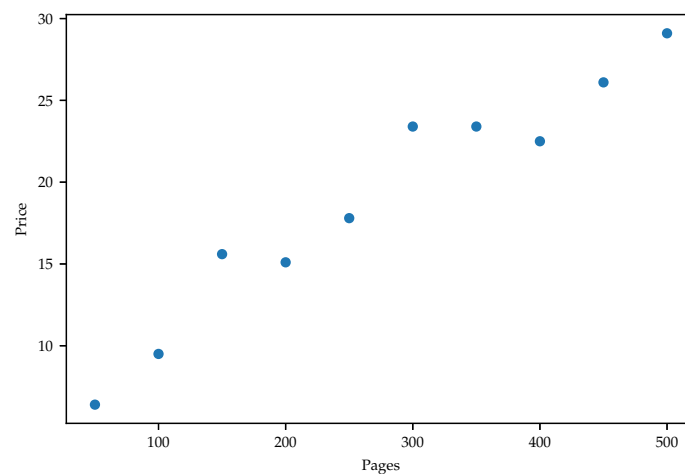


Figure 1.10: Scatter plot of the number of pages and book price.

Least Squares Method

Let us try to trace a line through *all* points in Figure 1.10 using a ruler. We can determine this is not possible (see Figure 1.11). Therefore, the points are only *approximately* on a straight line.

We then wonder: “How can we find a line that fits all points *as well as possible*?” This leads directly to the next question: What does “as well as possible” mean? To find a measure to decide how well the line fits brings up similar difficulties as with the determination of the variance.

We could choose the line such that when we add up the vertical differences between

Chapter 1 Descriptive Statistics

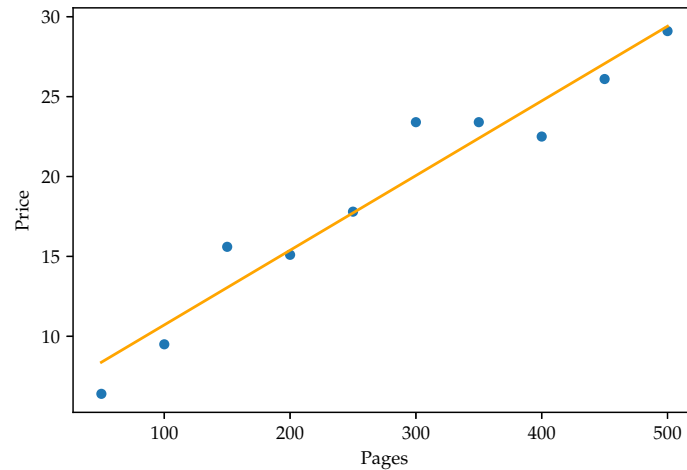


Figure 1.11: Line through the scatter plot

each observation and the line (see Figure 1.12), we assume a small sum of these distances indicates a good fit.

We designate the vertical difference between an observed point (x_i, y_i) and the line (the point on the line having coordinates $(x_i, a + bx_i)$) as the residual:

$$r_i = y_i - (a + bx_i) = y_i - a - bx_i$$

For our example, the residuals r_6 and r_8 for *this* line are represented in Figure 1.12. Residual r_6 is positive, since the point is above the line. Correspondingly, $r_8 < 0$.

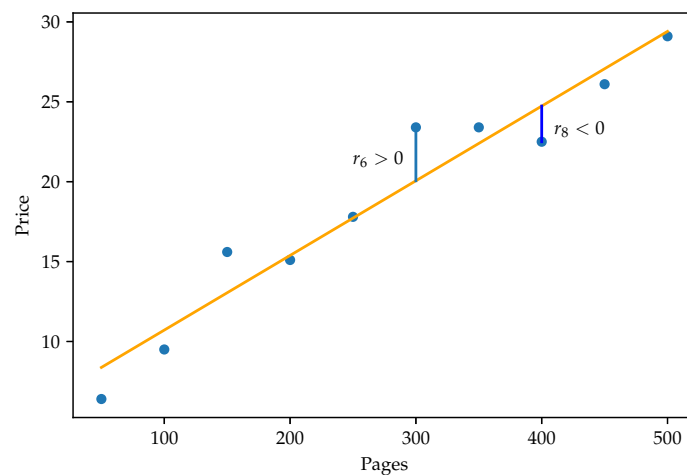


Figure 1.12: Residuals

We would like to determine the line $y = a + bx$ such that the sum

$$r_1 + r_2 + \dots + r_n = \sum_i r_i$$

is minimized. But this method has a serious weakness: If half the points are far above the line and the other half is far below it, then the sum of the deviations (residuals) can be close to zero, even though the line fits the data points very poorly. The positive deviations and the negative deviations simply cancel each other out. So, we must eliminate the deviations' signs before adding them. One option consists in adding up the absolute values, so $\sum_i |r_i|$, and minimizing this sum. However, since absolute values can be rather inconvenient to work with (for example, when trying to derive expressions with absolute values), so we turn to another possibility. It involves adding up the squares of the deviations, so

$$r_1^2 + r_2^2 + \cdots + r_n^2 = \sum_i r_i^2$$

The parameters a and b should be chosen so as to minimize the sum. The latter method is generally accepted, as it is much easier to work with than the absolute values. Hence a line optimally fits (according to our quality criteria) the points if the sum of the squares of the vertical deviations is minimized. This procedure is known as the least squares method. In our case, **Python** gives us the values $a = 6.04$ and $b = 0.047$.

Example 1.2.4

In our case, with **Python** we obtain the values $a = 6.04$ und $b = 0.047$.

```
b, a = np.polyfit(book["pages"], book["price"], deg=1)
print(a, b)
```

The linear equation is

$$y = 6.04 + 0.04673x$$

Therefore, the publisher's basic costs are about 6 Swiss francs. And it charges about 5 centimes per page. ◀

Remarks:

- i. The command

```
np.polyfit(book["pages"], book["price"], deg=1)
```

from **numpy** fits a polynomial of degree 1 (linear function) to the data. The output consists of two values: the first is the slope of the line, the second is the intercept.

- ii. This is also called the *regression line*. ♦

Example 1.2.5

Based on this model, how much would a 375-page book cost? We insert $x = 375$ in the linear equation above and obtain

$$y = 6.04 + 0.04673 \cdot 375 \approx 23.60$$

So, the book would cost about CHF 23.60. This model nevertheless only has limited validity. One has to be especially careful with *extrapolation*. We could calculate how much a book with a million pages would cost, but such amount certainly no longer matches reality.

The line in Figure 1.11 on page 28 is plotted in **Python** as follows:

```
book.plot(kind="scatter", x="pages", y="price")
b, a = np.polyfit(book["pages"], book["price"], deg=1)

x = np.linspace(book["pages"].min(), book["pages"].max())

plt.plot(x, a+b*x, c="orange")

plt.xlabel("Pages")
plt.ylabel("Price")

plt.show()
```

The command

```
x = np.linspace(book["pages"].min(), book["pages"].max())
```

generates a vector x of length 50. The first component of this vector is minimal value of pages in the Dataframe book; the last component is its maximal value. ◀

How does **Python** Calculate the Parameters a and b ?

Parameters a and b are determined as follows:

Parameters a and b should minimize the following expression (least squares method):

$$\sum_{i=1}^n (y_i - (a + bx_i))^2$$

The solution to this optimization problem yields

$$\hat{b} = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x}$$

where \bar{x} and \bar{y} are the corresponding averages. \hat{a} and \hat{b} are the estimations of parameters a and b , so the values for which $\sum_{i=1}^n (y_i - (a + bx_i))^2$ is the smallest.

Remarks:

- i. We are not establishing here how to calculate a and b . Just to give an idea: Since

$$\sum_i r_i^2 = \sum_{i=1}^n (y_i - (a + bx_i))^2$$

must be minimized, the derivative of $\sum_i r_i^2$ with respect to a and with respect to b must be equal to 0. So, we get a system of equations consisting of two equations and two unknowns:


$$\frac{\partial}{\partial a} \sum_{i=1}^n (y_i - (a + bx_i))^2 = \sum_{i=1}^n -2 (y_i - a - bx_i) \stackrel{!}{=} 0$$

$$\frac{\partial}{\partial b} \sum_{i=1}^n (y_i - (a + bx_i))^2 = \sum_{i=1}^n -2 (y_i - a - bx_i) \cdot x_i \stackrel{!}{=} 0$$

The algebraic manipulations leading to the estimations of a and b become rather long and are not derived here.

- ii. Here, \hat{a} and \hat{b} are always calculated with **Python**. ◆

Example 1.2.6

It can be conjectured that a relationship exists between the height of fathers and the height of their sons. In 1900, the British statistician Karl Pearson charted the height of 10 (in reality, it was 1078) randomly selected men against the height of their fathers. And he obtained the data in Table 1.4. 

There *appears* to be an actual relationship here: the taller the father, the taller the son. When we plot the scatter plot (see Figure 1.13), we see a (possible) linear relationship exists: The point cloud “follows” the straight line $y = 0.445x + 94.7$; these parameters were calculated from the data using the least squares method.

Chapter 1 Descriptive Statistics

So, for a father's height of 180 cm, which does not appear in Table 1.4, we can calculate the expected value of the son's height.

$$y = 0.445 \cdot 180 + 94.7 \approx 175 \text{ cm}$$

However, we must be careful not to use this formula where it is not applicable. For example, we obtain for $x = 0$ a value of 94.7. But what does that mean? If the father is 0 cm high, then the son would be about 95 cm high, according to this model. But that is nonsensical. ◀

Father's height	152	157	163	165	168	170	173	178	183	188
Son's height	162	166	168	166	170	170	171	173	178	178

Table 1.4: Height comparison of fathers and sons.

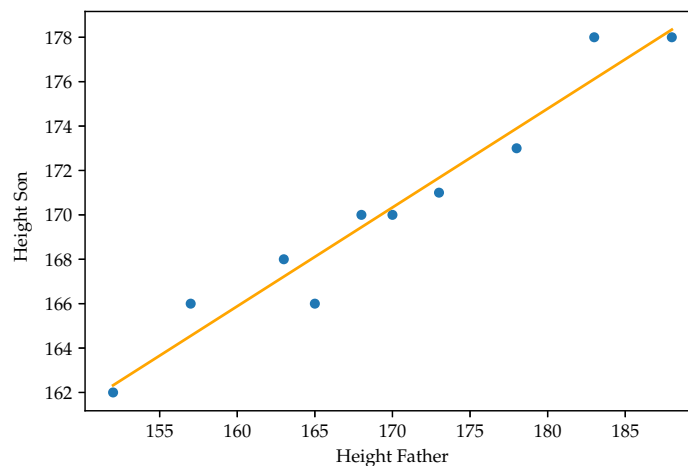


Figure 1.13: Scatter plot of father and son sizes

Example 1.2.7

The following table establishes a relationship between the numbers of traffic fatalities occurring in 1988 and 1989 in twelve counties (the counties exhibit more or less the same population size).

The table does not make a clear relationship evident. Let us consider the scatter plot in Figure 1.14. We can see there is no relationship. But this was also to be expected if we could reasonably assume there is no relationship between the traffic fatalities of different individual counties. Figure 1.14 charts the regression line. We can calculate and plot it. However, this makes not sense, since there is no linear relationship between the measurement quantities.

In the next subsection, we will study a quantity that allows us to make a declaration about how strong the linear relationship is between two measurement quantities. ◀

Chapter 1 Descriptive Statistics

County	1	2	3	4	5	6	7	8	9	10	11	12
1988 traffic fatalities	121	96	85	113	102	118	90	84	107	112	95	101
1989 traffic fatalities	104	91	101	110	117	108	96	102	114	96	88	106

Table 1.5: Traffic fatalities in two consecutive years.

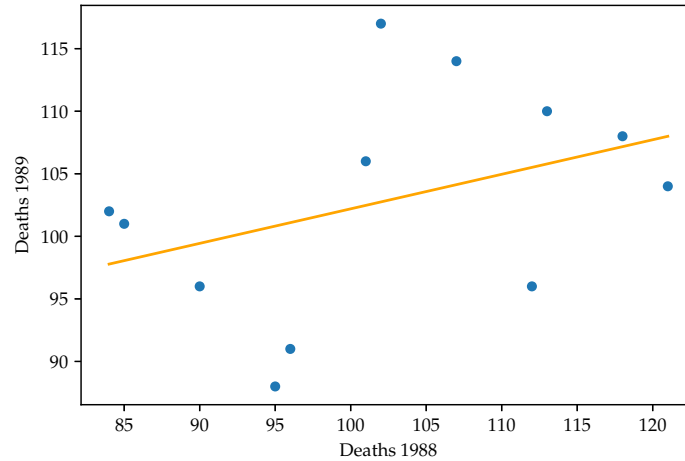


Figure 1.14: Traffic deaths

Example 1.2.8

As another example, let us consider again the survey of the relationship between wine consumption and the death rate. We will take as a basis for the data the linear model

$$y = a + bx$$

where x is the annual wine consumption per person and y is the mortality per 1000 people. We then can estimate the parameters a and b based on the data points, using the least squares method, and we obtain the regression line

$$y = 7.68655 - 0.07608x$$

Let us however consider the scatter plot with the regression line (see Figure 1.15); we determine the relationship between the measurement quantities is not linear. The scatter plot is more suggestive of a hyperbolic function.

So, the regression line says little here about the true relation. ◀

We can (almost) always determine the regression line. But in the last two examples, we saw that the regression line provides very little information about the actual distribution of points in the scatter plot. There are two reasons for this:

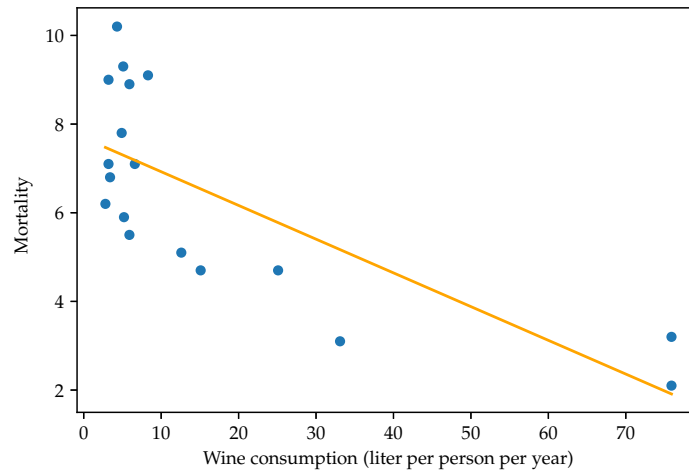


Figure 1.15: Regression line between wine consumption and death rate

- The points do not seem to follow any rule at all
- The points seem to follow a nonlinear rule

But how can we determine whether the data has a linear relationship or not? One possibility is sure to consider the situation graphically, as we just saw. But we can also give a value that describes the relationship numerically.

1.2.3 Empirical Covariance and Correlation

To quantify the linear dependence of two quantities, the **empirical correlation coefficient** r is the most common summary statistic (also written $\hat{\rho}$).

Empirical Covariance

We first introduce the notion of covariance.

Empirical Covariance

For the samples $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ the *empirical covariance* is defined as

$$\text{cov}_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

If $x = y$, then we have

$$\text{cov}_{xx} = \frac{\sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})}{n - 1}$$

and this is exactly the *empirical variance* s_x^2 of x :

$$\text{cov}_{xx} = \text{var}_x = s_x^2$$

Now we want to convince ourselves that indeed the covariance captures a linear relation.

Example 1.2.9

In Figure 1.16 you see a couple of data points which more or less follow a straight line. You also see the lines parallel to the coordinate axes $x = \bar{x}$ and $y = \bar{y}$.

For alignment, we subtract the mean value \bar{x} from the x coordinates and the mean value \bar{y} from the y coordinates. Then, we obtain Figure 1.17.

The empirical covariance for these points now reads

$$\text{cov}_{x^*y^*} = \frac{\sum_{i=1}^n x_i^* y_i^*}{n - 1}$$

Hence, in the numerator, all products $x_i^* y_i^*$ are summed up. In quadrants I and III, these products are positive, in quadrants II and IV they are negative.

Since in our example almost all points lie in quadrants I and III, $\text{cov}_{x^*y^*}$ becomes positive:

$$\text{cov}_{x^*y^*} > 0$$

If the points lie on a descending line, then the points are mostly in quadrants II and IV. The value of $\text{cov}_{x^*y^*}$ is then negative:

$$\text{cov}_{x^*y^*} < 0$$

Now what happens if the points show no linear relation (see Figure 1.18)?

In this case, the products $x_i^* y_i^*$ sum up to approximately zero, since roughly half of the points are in quadrants I and III (positive products), and roughly the other half lies in quadrants II and IV, being of similar magnitude in absolute value. Then, we have

$$\text{cov}_{x^*y^*} \approx 0$$

How does it look like for a quadratic relation (see Figure 1.19)?

Chapter 1 Descriptive Statistics

Also here, the products on the left and on the right of the y axis approximately cancel out. We have

$$\text{cov}_{x^*y^*} \approx 0$$

Note that this argument depends on the mirror symmetry of the parabola with respect to the y -axis. If there is a non mirror symmetric, non-linear (e.g., cubic) relationship, then the covariance does not have to sum up to zero. However, such relationships can also be interpreted as approximately linear.

The covariance essentially recognizes *linear* relationships. ◀

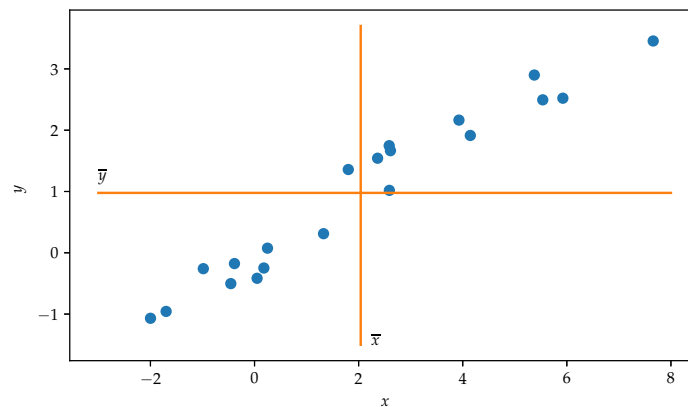


Figure 1.16: Points lying close to a straight line

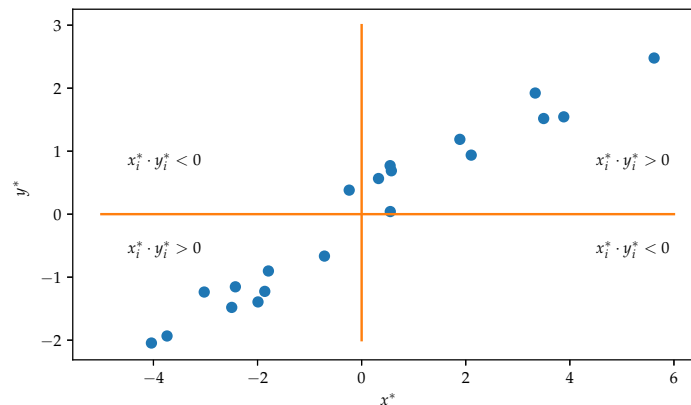


Figure 1.17: The mean values were subtracted from the coordinates.

When there is no linear relation, then the covariance can be 0, although there is a non-linear relationship. We should **never** rely only on the value of the covariance, instead we should **always** look at the plots to check for non-linear relationships.

Chapter 1 Descriptive Statistics

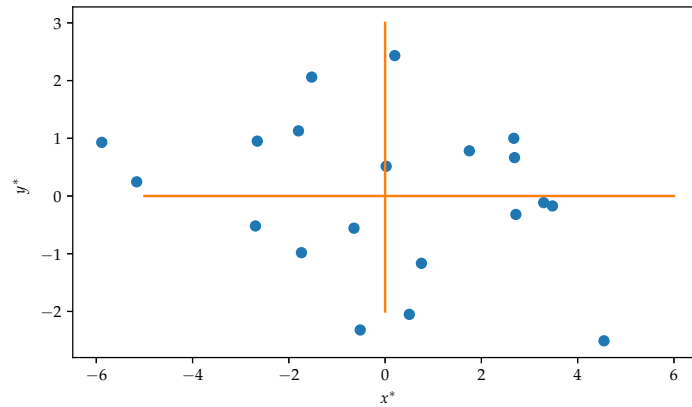


Figure 1.18: Points whose coordinates show no relation

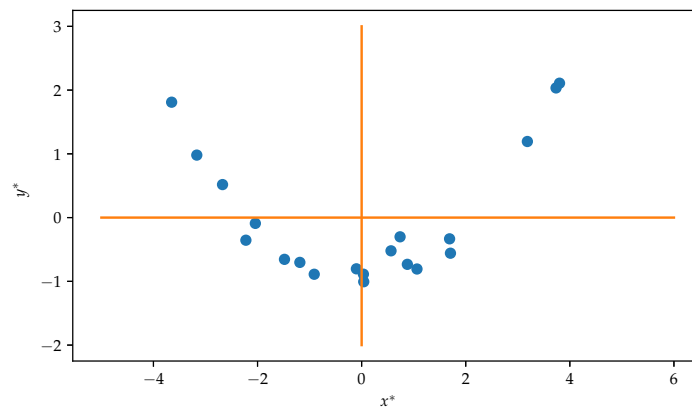


Figure 1.19: Points close to a parabola

Example 1.2.10

Benzopyrene is a cancerogene hydrocarbon molecule, which is the product of incomplete burnings. A source of Benzopyrene and carbon monoxide are car exhausts. Colucci and Begeman (1971) analyzed 16 air samples, which were taken at the Herald Square in Manhattan.

They recorded the carbon monoxide concentration (in parts per million) and the benzopyrene concentration (in microgram per 1000 cubic meter) for each sample. In the file **Herald.dat**, the data are listed:

```
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
from pandas import Series, DataFrame
df = pd.read_table("Herald.dat")
print(df.head())
```

Now we represent the data graphically in Figure 1.20.

```

1 import matplotlib.pyplot as plt
2 import numpy as np
3 import pandas as pd
4 from pandas import Series, DataFrame
5
6 df = pd.read_table("Herald.dat")
7
8 df.plot(kind="scatter", x="CO", y="Benzoa")
9
10 plt.plot((2.5, 20), (df["Benzoa"].mean(), df["Benzoa"].mean()),
11          c="orange")
12
13 plt.plot((df["CO"].mean(), df["CO"].mean()), (0, 10), c="orange")
14
15 plt.ylabel("Benzoapyrene")
16 plt.show()

```

With **Python** we can now calculate the empirical covariance of the Herald Square pairs.

```
df.cov()
```

The Output of **Python** is a so-called *covariance matrix*. The only value important to us is in row **CO** and column **Benzoa** (or vice versa). Thus, we have

$$\text{COV}_{\text{df}["\text{CO}"]\text{df}["\text{Benzoa}]} = 5.51$$

The value 25.8 in the covariance matrix corresponds to the variance of **CO**. Accordingly, the value 9.33 is the variance of **Benzoa**. ◀

The problem with the covariance is that it is hard to be interpreted. How far from zero does it have to be in order to indicate a relationship? Also, the covariance in Example 1.2.10 has the unit

$$\text{ppm} \cdot \mu\text{g}/1000\text{m}^3$$

which is almost impossible to interpret.

In order to remedy this, we introduce the *empirical correlation coefficient* as a new normalized and dimensionless quantity to measure a linear relationship between sample pairs (x_i, y_i) .

Empirical Correlation Coefficient

The most used measure for the linear relationship between two quantities is the *empirical correlation coefficient* or *Pearson correlation coefficient* r_{xy} (also denoted with r or $\hat{\rho}$).

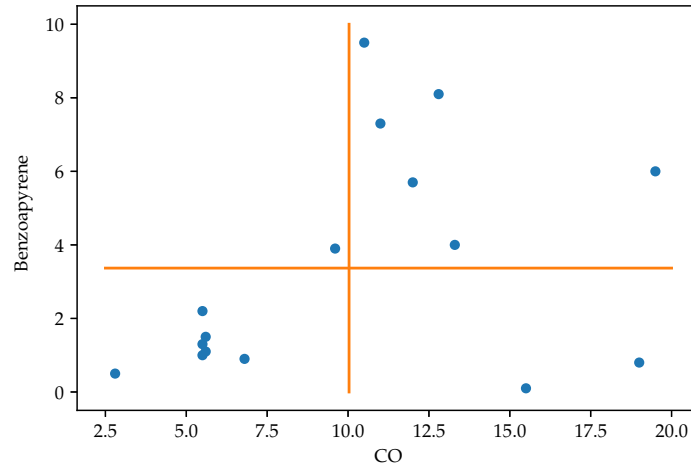


Figure 1.20: Scatter diagram of benzo(a)pyrene versus carbon monoxide

It is defined by standardization of the covariance. This means that the covariance is divided by the standard deviations of x and that of y . Then the correlation coefficient takes values between -1 and 1 , where a value of 0 indicates **no linear** relationship.

Empirical Correlation Coefficient

The empirical correlation coefficient r for the coordinate pairs (x_i, y_i) is defined as follows:

$$r_{xy} = \frac{\text{cov}_{xy}}{s_x \cdot s_y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{(n-1) \cdot s_x \cdot s_y}$$

where s_x and s_y are the empirical standard deviations of the samples x_i and y_i .

The empirical correlation coefficient is a dimensionless number between -1 and $+1$ that measures the strength and direction of the *linear dependence* between the x and y data. The empirical correlation has the following properties:

1. If $r_{xy} = +1$, then the points lie on an increasing straight line ($y = a + bx$ with $a \in \mathbb{R}$ and $b > 0$) and vice versa.
2. If $r_{xy} = -1$, then the points lie on a decreasing straight line ($y = a + bx$ with $a \in \mathbb{R}$ and $b < 0$) and vice versa.
3. If x and y are independent of each other (i.e., there is no relationship), then $r_{xy} \approx 0$.

The converse is not generally true: $r = 0$ does *not* mean x and y are independent of each other (see Figure 1.21 on page 41)

One direction of the first two properties are easy to understand: if we substitute $y_i = a + bx_i$ and $\bar{y} = a + b\bar{x}$ into the expression for the correlation coefficient, then we obtain

$$\begin{aligned} r &= \frac{\sum_{i=1}^n (x_i - \bar{x})(bx_i + a - (b\bar{x} + a))}{\sqrt{(\sum_{i=1}^n (x_i - \bar{x})^2) \cdot (\sum_{i=1}^n (bx_i + a - (b\bar{x} + a))^2)}} \\ &= \frac{b \cdot \sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})}{|b| \cdot \sqrt{(\sum_{i=1}^n (x_i - \bar{x})^2) \cdot (\sum_{i=1}^n (x_i - \bar{x})^2)}} \\ &= \text{sign}(b) \end{aligned}$$

where the sign of a number is defined as

$$\text{sign}(b) = \begin{cases} 1 & \text{if } b > 0 \\ 0 & \text{if } b = 0 \\ -1 & \text{if } b < 0 \end{cases}$$

However, r should never be calculated without paying attention to the scatter plot, since the same value of r_{xy} can result from very different structures. See Figure 1.21 on the following page.

Example 1.2.11

In our example about page numbers and prices of books, **Python** gives us

```
book.corr().iloc[0,1]
```

the value 0.968. So, it is very close to 1, hence there is a strong linear relationship. Furthermore, the correlation coefficient is positive, hence the slope of the regression line is positive, which corresponds to a “the more, the more”, thus a positive linear relationship. ◀

Remark 1.2.1

The command

```
book.corr()
```

is more general and produces the so-called correlation matrix.

Example 1.2.12

Also in the example of the heights of fathers and sons, we expect a high correlation coefficient. We obtain 0.973. ◀

Example 1.2.13

With the traffic accidents, there is no relationship, and we expect a correlation coefficient close to zero. It is 0.386. ◀

Example 1.2.14

In the wine consumption case, we expect a negative correlation coefficient, since with increasing wine consumption we have decreasing mortality. It is -0.746 . Without representing the data in the scatter plot, this value would erroneously lead to the conclusion there is a strong negative linear relationship. ▶

Example 1.2.15

Figure 1.21 represents 21 different data sets, consisting of as many pairs of observations (x_i, y_i) with the corresponding points in the scatter plot. The associated empirical correlation appears over each data set. ▶

A perfectly linear relationship has the empirical correlation $+1$ or -1 (depending on whether the slope is positive or negative; see the second row in Figure 1.21). The more the points are scattered around the linear relationship, the small the empirical correlation's magnitude (see the first row).

Since the empirical correlation only measures the *linear* relationship, there can be a (nonlinear) relationship between the two variables x and y even if the empirical correlation is zero (see the bottom row in Figure 1.21). ▶

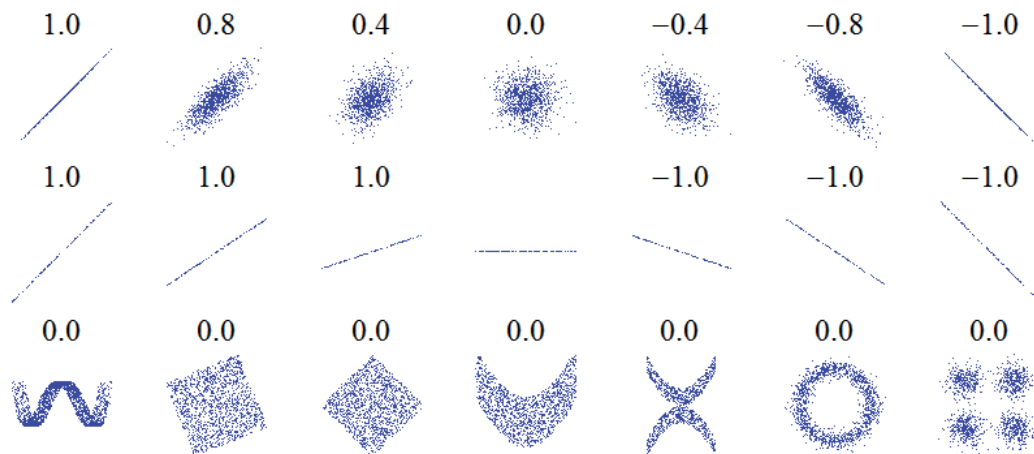


Figure 1.21: 21 different data sets and their empirical correlation coefficients.

Example 1.2.16

For the dataset **Herald** the correlation coefficient is calculated as follows

```
df.corr()
```

We obtain the correlation matrix. Again, the entry in row **CO** and column **Benzoa** is essential. We have

$$r_{df["CO"],df["Benzoa"]} = 0.3551$$

Although the correlation is only moderate, there is a physical explanation for a correlation, since both chemicals are the product of incomplete combustion. Typically, a correlation of ≈ 0.3 visually yields a slight impression that y increases if x is increased. However, the points scatter rather strongly. ◀

Educational Objectives

You are able to ...

- ☐ explain and interpret the most important methods of descriptive statistics
- ☐ calculate the following quantities: arithmetic mean, standard deviation, variance, quantile, median, covariance, and correlation coefficient.
- ☐ explain the basic idea of simple linear regression; to define the regression model, to interpret the coefficients, and to explain how these are estimated.
- ☐ explain the construction of the following graphical representations: histogram, box plot, empirical cumulative distribution function, and scatter plot.

Computer-based Educational Objectives

You are able to ...

- ☐ read data from a file with the help of **pandas** as a **Series** or **DataFrame**.
- ☐ calculate the mean and the empirical standard deviation using the **pandas**-methods **var()**, resp. **std()**.
- ☐ find the empirical median and the quantiles with the **pandas**-methods **median()** and **quantile** and from there the interquartile range.
- ☐ plot a histogram with the **pandas**-method **plot(kind="hist")**.
- ☐ produce a boxplot for some dataset using the **pandas**-method **plot(kind="box")**.

Chapter 1 Descriptive Statistics

- ☐ create a scatter plot for a two-dimensional data set with the help of the **pandas**-method `plot(kind='scatter')`.
- ☐ determine the coefficients of a simple linear regression model using the **numpy**-method `np.polyfit()`.
- ☐ calculate the empirical covariance and the correlation coefficienten for two quantities with the **pandas**-methods `cov()` and `corr()`.