The Elements of Multi-Variate Analysis for Data Science

Mohammad Samy BALADRAM* and Nobuaki OBATA

Graduate School of Information Sciences, Tohoku University, Sendai 980-8579, Japan

These lecture notes provide a quick review of basic concepts in statistical analysis and probability theory for data science. We survey general description of single- and multi-variate data, and derive regression models by means of the method of least squares. As theoretical backgrounds we provide basic knowledge of probability theory which is indispensable for further study of mathematical statistics and probability models. We show that the regression line for a multi-variate normal distribution coincides with the regression curve defined through the conditional density function. In Appendix matrix operations are quickly reviewed. These notes are based on the lectures delivered in Graduate Program in Data Science (GP-DS) and Data Sciences Program (DSP) at Tohoku University in 2018–2020.

KEYWORDS: data matrix, method of least squares, multi-variate analysis, regression analysis, probability distribution

1. Data and Statistical Analysis

1.1 Data matrices

A set of characteristics collected from objects is called data in general. The totality of objects to be measured or surveyed is called a population and each member therein an individual. Thus, the data are collected from each individual in a target population. Data consisting of values obtained by measuring an amount are called quantitative data. An example is shown in Table 1.1, which is the list of height, weight and age of the players of Team A. In this survey the set of all players of Team A is a population and each player is an individual. The full list is deferred to Appendix for the readers’ exercise.

With each survey item we associate a variable or a variate, which means a measurable quantity varying a certain range of real numbers. The term “variate” is often used in the context of physical, economical or statistical surveys, while the term “variable” is very common in any context of mathematics. Thus, data are a collection of values of the variable corresponding to a measurement. For example, for the data in Table 1.1 one may associate a variable \( x \) to height, \( y \) to weight and \( z \) to age. Then the values of data of the \( i \)th individual are denoted by

\[
\begin{align*}
\tilde{x}_1, y_1, z_1.
\end{align*}
\]

In this fashion we have \( x_1 = 178, y_2 = 90 \) and \( z_{82} = 24 \). When the number of variables is large, instead of assigning different symbols such as \( x, y, z, \ldots \) to variables, we use a symbol with indices. For example, we may assign \( x_1 \) to height, \( x_2 \) to weight and \( x_3 \) to age. Note that \( x_1, x_2, x_3 \) are not the data of the first three individuals, but the three variables. In this case the \( i \)th data are denoted by

\[
\begin{align*}
\tilde{x}_i, y_i, z_i.
\end{align*}
\]
At first glance those notations appear confusing, but after some practice the usefulness will be understood.

Let $x_1, \ldots, x_j, \ldots, x_p$ denote the variables corresponding to $p$ measurements. After surveying $n$ individuals we obtain $p$-variate data or $p$-dimensional data, where the values of data of the $i$th individual are denoted by

$$x_{i1}, x_{i2}, \ldots, x_{ip}.$$ 

In other words, $x_{ij}$ denotes the value of the variable $x_j$ of the $i$th individual. The data matrix is an $n \times p$ matrix with $x_{ij}$ being an $(i, j)$ entry as follows:

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nj} & \cdots & x_{np} \end{bmatrix}$$

The number of rows coincides with that of individuals and the number of columns coincides with that of variables. For example, the data matrix obtained from Table 1.1 becomes an $83 \times 3$ matrix. In these lecture notes we will be concerned only with quantitative data given in the form of data matrix, for which the very powerful mathematical tools of linear algebra are available effectively, see Sect. 3.

Remark 1.1. There are other types of data called qualitative data, which are recorded in terms of letters, symbols, diagrams and so on. If the individuals are classified into categories by some nominal attribute, we obtain nominal data or category data. For example, the nominal attribute “sex” gives rise to two categories “male” and “female.” Likewise, “nationality” gives rise to quite a few categories such as American, English, French, Japanese, ..., . These data are often quantified by using dummy variables for further analysis. For example, two categories “male” and “female” are represented by 0 and 1, respectively. Another type of qualitative data is ordinal data. The results of a questionnaire survey of customer satisfaction are recorded in terms of a few grades, such as A, B and C for three grades. Even if these grades might be recorded in terms of numbers such as 1, 2 and 3, those numbers indicate only the grades or the order. Hence the difference or the ratio among those numbers do not make any sense in general.

1.2 Statistical analysis

Such data as shown in Table 1.1 or in the form of data matrix (1.1) are called raw data in the sense that the data are in original form, collected directly from observation, unorganized and uncooked. The raw data are usually entered into a computer system in a suitable form according to a software. A common spreadsheet software requires the input form just as in Table 1.1.

A data matrix being just a large array of values, it is difficult to extract useful information at a glance. What we need is reduction of data. Given a data matrix $X$, we apply some functions $f$ to get new values called statistics, which clarify characteristics of data. If $X$ consists of $np$ values as shown in (1.1), a function of $X$ is in fact a function of $np$ variables. The main purpose of these lecture notes is to study basic statistics and their applications.

Upon closing this introductory section we mention a few remarks on statistical inference. In statistical analysis it is essential to distinguish a target population and surveyed individuals. A survey that measures the entire population is called a complete survey or a census. A national population census is a typical example. It would be, however, impractical to perform a complete survey for reasons of size, time, cost and so forth. In most cases we select some individuals from a target population, each of which is called a sample. A survey that measures only selected samples is called a sample survey. Examples include an audience rating survey, a public-opinion poll, a sampling inspection of products and so forth. Moreover, usual experiments or observations are in principle regarded as sample surveys. A sample survey saves the cost and time, and makes it easier to maintain high quality information, whereas it can not get away from sampling errors because it measures only a part of a target population. In this context statistical inference becomes essential for estimating population characteristics from sample data and it is the main theme of mathematical statistics.

Data collected over time are called time series data, where data are listed in time order. In the form of a data matrix $X = [x_{ij}]$ we understand $i$ as a time parameter. Examples include counts of sunspots, weather data, stock data, traffic accident outbreaks and so forth. In this context prediction becomes a main theme, where probability models (stochastic processes) play an essential role. Interested readers should refer to suitable books for further study.
2. Summarizing Single-Variate Data

2.1 Frequency table and histogram

Consider a single variable \( x \) and suppose we are given single-variate data of size \( n \) as \( x_1, x_2, \ldots, x_n \) (2.1).

According to the standard notation of data matrix introduced in Sect. 1, the above data (2.1) should be written in the form of a column vector. However, saving space is in priority here as there is no danger of confusion.

In practice, a single-variate data is a long sequence of numbers and we are not able to find useful information at a glance. The first task is to classify the data and extract information concerning how the data are distributed on the real line \( \mathbb{R} \) or on the \( x \)-axis. Take an interval \( I \subset \mathbb{R} \) containing all the data and divide \( I \) into a few small intervals of equal width:

\[
I : c_0 < c_1 < \cdots < c_k.
\]

Each small interval \( I_i = [c_{i-1}, c_i) \) is called a class. The midpoint of \( I_i = [c_{i-1}, c_i) \) defined by

\[
a_i = \frac{c_{i-1} + c_i}{2}
\]

is called the class mark. A class mark is used to represent the values in the interval \( I_i \).

![Fig. 2.1. Classification of data.](image)

Each value of the data (2.1) falls into a unique class \( I_i = [c_{i-1}, c_i) \). Then for each class \( I_i \) we count the number of values falling into it, which is referred to as the (absolute) frequency. If \( f_i \) is the frequency for \( I_i \), the ratio

\[
p_i = \frac{f_i}{n}
\]

is called the relative frequency, where \( n \) is the total number of the data. Finally the results are summarized in the form of frequency table as shown in Table 2.1.

| Classes \( I_i \) | Class marks \( a_i \) | Frequency \( f_i \) | Relative frequency \( p_i \) |
|------------------|----------------------|-----------------|-----------------|
| \( I_1 \)        | \( a_1 \)             | \( f_1 \)       | \( p_1 \)       |
| \( I_2 \)        | \( a_2 \)             | \( f_2 \)       | \( p_2 \)       |
| \( \vdots \)     | \( \vdots \)          | \( \vdots \)    | \( \vdots \)    |
| \( I_k \)        | \( a_k \)             | \( f_k \)       | \( p_k \)       |
| Total            |                      | \( n \)         | 1               |

There is no strict rule to decide the number of classes or their width. As the width of a class becomes wider, we lose more information on distribution of the data. Conversely, as the width becomes narrower, the outline of the distribution is difficult to grasp. It is recommended to make some trials.

Graphical representation of a frequency table is useful. On each small interval \( I_i \) in the \( x \)-axis we draw a rectangle with height proportional to the frequency \( f_i \) or equivalently the relative frequency \( p_i \). These rectangles are not separated since the \( x \)-axis stands for a continuous scale. The diagram obtained in this way is called a histogram. The graph obtained by connecting the midpoints of the tops of the histogram by straight lines is called a frequency polygon.

Another useful statistic is a cumulative frequency. For each class \( I_i \) the cumulative frequency is defined by

\[
f_1 + f_2 + \cdots + f_i.
\]

Likewise we define cumulative relative frequency, which is also called cumulative percentage. We will see in Sect. 4 that the cumulative relative frequency is a bridge connecting probability theory and practical statistical analysis.
Example 2.1. The frequency table of height of players of Team A is shown in Table 2.2, where the cumulative frequencies and the relative cumulative frequencies are added. The left diagram in Fig. 2.2 shows the histogram together with a frequency polygon. The right diagram in Fig. 2.2 shows the cumulative relative frequencies.

2.2 Measures of centrality

Suppose we are given single-variate data of size \( n \) for a variable \( x \) as in (2.1). We now look for a suitable value that represents a center of the data. Most commonly used is the mean or the average defined by

\[
\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i,
\]

that is, by adding up all the values and dividing by the number of data. As there are many variants of “mean” in mathematics, to avoid confusion the mean defined by (2.2) is called the arithmetic mean. Instead of raw data, we may start with a frequency table given as in Table 2.1, where \( f_i \) is a frequency of a class \( I_i \) with class mark \( a_i \). Then the mean is defined by

\[
\bar{x} = \frac{1}{n} \sum_{i=1}^{k} a_i f_i.
\]

Using the relative frequency \( p_i = f_i / n \), we obtain a useful formula:

\[
\bar{x} = \sum_{i=1}^{k} a_i p_i.
\]

We will notice in Sect. 5 that (2.4) is consistent with the definition of mean (or expected value) of a random variable.

It is noted that a frequency table sacrifices some information of the original raw data. Once the raw data are transferred into a frequency table, the exact values of data are not recovered. From a frequency \( f_i \) of a class \( I_i = [c_{i-1}, c_i] \) we only know that there are \( f_i \) values in the raw data lying in the interval \( I_i = [c_{i-1}, c_i] \). We then understand that those values are equally distributed across the interval. In fact, the formula (2.3) or (2.4) is based on this interpretation. As a result, the mean computed directly from raw data and that from a frequency table do not coincide.

Exercise 2.2. Given single-variate data let \( \bar{x} \) be the mean of the original raw data and \( a \) the mean calculated from a frequency table summarizing the same data with classes of width \( d \). Show that

\[
\bar{x} = \frac{1}{n} \sum_{i=1}^{k} a_i f_i = \sum_{i=1}^{k} a_i p_i.
\]

\[
\bar{x} = \frac{1}{n} \sum_{i=1}^{k} a_i f_i = \sum_{i=1}^{k} a_i p_i.
\]
Rearranging the data (2.1) from smallest to largest in such a way that
\[ x_1 \leq x_2 \leq \cdots \leq x_n, \tag{2.5} \]
we call \( x_i \) the \( i \)th order statistic. In particular, the minimum of data is defined by
\[ \min x = \min\{x_1, x_2, \ldots, x_n\} = x_1, \tag{2.6} \]
and the maximum by
\[ \max x = \max\{x_1, x_2, \ldots, x_n\} = x_n. \tag{2.7} \]
The value at the middle position in (2.5) is called the median. If we have an odd number of data, the median is exact because the middle rank is determined among \( n \) data. If we have an even number, the median is defined to be the average of two data at the middle rank. To be precise, the median is defined by
\[
\text{med } x = \text{med}\{x_1, x_2, \ldots, x_n\} = \begin{cases} 
\frac{x_{\lfloor n/2 \rfloor}}{2} & \text{if } n \text{ is odd,} \\
\frac{1}{2} (x_{\lfloor n/2 \rfloor} + x_{\lfloor n/2 floor + 1}) & \text{if } n \text{ is even.}
\end{cases}
\]
Another candidate of representing data is the mode, defined to be the most frequently occurring value among the data. Usually the mode is applied to the frequency table, where the mode appears as a peak of the histogram.

The three statistics, mean, median and mode are most commonly used for central values of data. It is noted that there is no relation among the three statistics. In fact, any order of the three occurs as is easily seen by simple extremal examples.

**Example 2.3.** Figure 2.3 is the histogram of annual family income in Japan in 2016,* where the horizontal axis shows annual income in ten thousand yen and the vertical one the relative frequencies. Note that the values of 2000 or above are bundled into just one class. A significant feature of the histogram, commonly observed in similar surveys, is that the distribution spreads along a one-sided long tail (that is why the values of 2000 or above are bundled into one class). The mean, median and mode are given by
\[
\text{mean} = 560, \quad \text{median} = 442, \quad \text{mode} = 350.
\]
Which value to use for representing the center of data depends on purposes.

![Fig. 2.3. Annual family income in Japan in 2016.](image)

**Example 2.4.** Demography is an interesting research topic. Figure 2.4 is the histogram of the Japanese population by age in 2018,† where the horizontal axis shows the age in year and the vertical axis the population in ten thousand. Note that the ages of 100 or above are bundled into just one class. We know that
\[
\text{mean} = 47.2, \quad \text{median} = 47, \quad \text{mode} = 69.5.
\]

---

*Source: Comprehensive Survey of Living Conditions, Ministry of Health, Labour and Welfare, Japan, 2017.

†Source: Population Estimates, Portal Site of Official Statistics of Japan (e-Stat).
The mean and median are almost in coincidence, while the mode is fairly larger. Moreover, we find a significant feature that the histogram shows the second peak at the age of 45.5.

**Remark 2.5.** The definition of mode adopted in these lecture notes is based on the traditional descriptive statistics. If the highest frequency appears at two or more classes, the mode is not uniquely defined. From the histogram in Fig. 2.4 one may expect some significant meaning of peaks of the histogram. In some literature the term “mode” is used for any class mark that attains a peak. The latter definition is more common in theoretical study of the shape of distributions.

**2.3 Measures of variability**

In the previous subsection we introduced statistics that present the centrality of data. However, many different sets of data could have the same centrality. The next key to characterize distributions of data is to observe variability of data.

For given data \(x_1, x_2, \ldots, x_n\) of size \(n\) let

\[x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}\]  

(2.8)

be the rearrangement from smallest to largest. The minimum, maximum and median are already introduced as order statistics. Along a similar line we define the *first quartile* to be the value at the first quarter position in (2.8). Likewise the value at the third quarter position is called the *third quartile*. The former is denoted by \(Q_1\) and the latter by \(Q_3\). (The precise definition of these quartiles will be mentioned at the end of this subsection.) The set of the five statistics

\[
\text{min}, \ Q_1, \ \text{med}, \ Q_3, \ \text{max}
\]

is called the *five-number summary* of Tukey. A *box plot* is often used for its graphical representation, see Fig. 2.5. Occasionally in some literatures, the median in the box plot is replaced with the mean.

![Box plot](image)

**Fig. 2.5.** Box plot.

A simple index for variability of data is given by the *range*, which is by definition the difference between the maximum and minimum:

\[R = \max - \min.\]

It is noted, however, that the range is affected heavily by extremal values in the data. In that sense the difference between the first and third quartiles
IQR = \( Q_3 - Q_1 \),
called the interquartile range, is more useful for variability of the data. The interquartile range corresponds to the length of the box part of a box plot.

From both practical and theoretical aspects, a better statistic for variability of data is variance. Let \( x_1, x_2, \ldots, x_n \) be \( n \) data with mean \( \bar{x} \). Then the variance of the data is defined by

\[
s^2 = s^2_x = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2,
\]

which is the average of the squared deviation of each \( x_i \) from the mean \( \bar{x} \). When we need to clarify the variable \( x \) we write \( s^2_{x_i} \) but when there is no danger of confusion we write \( s^2 \) just for simplicity. Apparently, \( s^2 \geq 0 \) by definition. The positive root of the variance:

\[
s = s_x = \sqrt{s^2_x} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2}
\]
is called the standard deviation.

Expanding \((x_i - \bar{x})^2\) in the right-hand side of (2.9), we obtain

\[
s^2_x = \frac{1}{n} \sum_{i=1}^{n} (x_i^2 - 2\bar{x}x_i + \bar{x}^2) = \frac{1}{n} \sum_{i=1}^{n} x_i^2 - 2\bar{x} \frac{1}{n} \sum_{i=1}^{n} x_i + \frac{1}{n} \sum_{i=1}^{n} \bar{x}^2.
\]

The second term becomes \(-2\bar{x}^2\) by definition of the mean. The third term stands for the average of a constant value \( \bar{x}^2 \) independent of \( i \) and is equal to \( \bar{x}^2 \). Thus, (2.10) becomes

\[
s^2_x = \frac{1}{n} \sum_{i=1}^{n} x_i^2 - 2\bar{x}^2 + \bar{x}^2 = \frac{1}{n} \sum_{i=1}^{n} x_i^2 - \bar{x}^2.
\]

Recall that the bar notation \( \bar{x} \) stands for the mean of the variable \( x \). Accordingly, the mean of the variable \( x^2 \) is denoted by \( \bar{x^2} \). Thus, (2.11) is written in a concise form:

\[
s^2_x = \bar{x^2} - \bar{x}^2,
\]

where the right-hand side is equal to the mean of the square of \( x \) minus the square of the mean of \( x \).

**Example 2.6.** Table 2.3 lists basic statistics of height of players of Team A and the box plots is shown in Fig. 2.6.

| Table 2.3. Basic statistics: Height of players of Team A. |
|---------------------------------|-----|
| Size of data (n)                | 83  |
| Mean (\( \bar{x} \))           | 179.8 |
| Minimum (min)                  | 168.0 |
| First quartile (\( Q_1 \))     | 175.5 |
| Median (med)                   | 180.0 |
| Third quartile (\( Q_3 \))     | 183.5 |
| Maximum (max)                  | 196.0 |
| Range (\( R \))                | 28.0 |
| Interquartile range (IQR)      | 8.0  |
| Variance (\( s^2 \))           | 28.82 |
| Standard deviation (s)         | 5.37 |

\[
\text{Fig. 2.6. Box plot: Height of players of Team A.}
\]
Remark 2.7. There is an important variant of variance. Recall the definition of variance (2.9), where the sum of squared deviations is divided by \( n \). Instead of dividing by \( n \), we define a new statistic by
\[
\sigma^2 = \sigma_i^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2,
\]
(2.13)
which is called the unbiased variance. The difference is small but crucial. To draw a clear line the variance defined by (2.9) is called sample variance. However, the use of these terminologies is mixed up in literatures and we need to take care. In the Excel commands “VAR.P” is for sample variance and “VAR.S” for the unbiased variance.

Exercise 2.8. For data \( x_1, x_2, \ldots, x_n \) prove that \( s_x^2 = 0 \) occurs if and only if the data are constant.

Exercise 2.9. Let \( x_1, x_2, \ldots, x_n \) be data of size \( n \). Find the value \( a \) that minimizes the sum of squared deviation from \( a \):
\[
\sum_{i=1}^{n} (x_i - a)^2.
\]

Exercise 2.10. Let \( x_1, x_2, \ldots, x_n \) be data of size \( n \). Find the value \( a \) that minimizes the sum of modulus of deviation from \( a \):
\[
\sum_{i=1}^{n} |x_i - a|.
\]

Generalizing the quartiles, we use percentiles to report relative standing of an individual within a given data set. Roughly speaking, the 85th percentile is the value that is greater than or equal to 85% of all the values and less than or equal to the remaining 15%. Of course, this definition is not strict because the 85th percentile is not determined only by the above condition. In practice, for the \( k \)th percentile of a given data of size \( n \) we apply the following steps:

Step 1) Order all the values in the data set from smallest to largest, say,
\[
x_1 \leq x_2 \leq \cdots \leq x_{n-1} \leq x_n.
\]

Step 2) Calculate \( r = nk/100 \).

Step 3) If \( r \) is an integer, count the number in the ordered data (2.14) from left to right until we reach \( r \). Then the \( k \)th percentile is defined to be the average of \( x_{(r)} \) and \( x_{(r+1)} \).

Step 4) If \( r \) is not an integer, round it up to the nearest integer to obtain \( s = \lfloor r \rfloor \). Then, count the number in the ordered data from left to right until we reach \( s \). Then the \( k \)th percentile is defined to be the value of \( x_s \).

As is easily seen, the median coincides with the 50th percentile. The first and third quartiles are defined to be the 25th and 75th percentiles, respectively.

Example 2.11. Below is a list of 25 test scores ordered from lowest to highest:

| 43 | 54 | 56 | 61 | 62 | 66 | 68 | 69 | 69 | 70 | 71 | 72 | 77 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 78 | 79 | 85 | 87 | 88 | 89 | 93 | 95 | 96 | 98 | 99 | 99 |

Let us find the 90th percentile. Multiplying 90% with the total number of scores, we obtain 0.9 \( \times 25 = 22.5 \). This is not an integer. Rounding it up to the nearest integer, we obtain 23. Counting the number of ordered data from left to right, we find the 23rd value in the data. That is 98, which is the 90th percentile of the given data. For the 20th percentile, first take \( 0.20 \times 25 = 5 \). This is an integer so the 20th percentile is the average of the 5th and 6th values in the ordered data. Thus the 20th percentile is given by \( \frac{62 + 66}{2} = 64 \).

Remark 2.12. Our definition of a percentile is just for practical use and is less theoretical. The idea of percentile is more suitable to a continuous distribution function (or a density function), and plays an essential role in statistical estimation and hypothesis testing.

2.4 Normalization

Let \( x_1, x_2, \ldots, x_n \) be data of the variable \( x \). Recall that the mean and variance are defined by
\[
\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i, \quad s_x^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2,
\]
respectively. The standard deviation \( s_x \) is by definition the positive square root of the variance. The new variable
\[
\tilde{x} = \frac{x - \bar{x}}{s_x}
\]
(2.15)
is called the normalization of \( x \).
Theorem 2.13. Given data \( x_1, x_2, \ldots, x_n \), let \( \bar{x}_1, \bar{x}_2, \ldots, \bar{x}_n \) be their normalization. Then the normalized data has mean 0, variance 1, and hence standard deviation 1. That is, 
\[
\bar{x} = 0, \quad s_x^2 = 1, \quad s_{\bar{x}} = 1.
\]

Proof. By definition of normalization (2.15) we have
\[
\bar{x} = \frac{1}{n} \sum_{i=1}^{n} \frac{x_i}{\bar{x}},
\]
and after simple algebra we come to
\[
\bar{x} - \bar{x} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}) = \frac{1}{n} \left( \frac{1}{n} \sum_{i=1}^{n} x_i - \frac{1}{n} \sum_{i=1}^{n} \bar{x} \right) = \frac{1}{n} (\bar{x} - \bar{x}) = 0.
\]
Then the variance of the normalized data is given by
\[
s_{\bar{x}}^2 = \frac{1}{n} \sum_{i=1}^{n} (\bar{x} - \bar{x})^2 = \frac{1}{n} \sum_{i=1}^{n} \bar{x}_i^2.
\]
Again by definition of normalization (2.15) we come to
\[
s_{\bar{x}}^2 = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{x_i - \bar{x}}{s_x} \right)^2 = \frac{1}{s_x^2} \sum_{i=1}^{n} \left( x_i - \bar{x} \right)^2 = \frac{1}{s_x^2} \times s_x^2 = 1
\]
as desired. \( \square \)

There are several merits of normalization of data. As a rule, original data are real numbers with certain unit associated to a measurement, and hence the values depend on the choice of the unit as well as the origin of a scale. After normalization the effect of such freedom is cancelled. In fact, the normalization (2.15) depends only on the difference between \( x_i \) and the mean \( \bar{x} \), moreover, it is free from the unit after taking the ratio to the standard deviation. For example, direct comparison of two measured values 172 cm (height) and 65 kg (weight) does not make a sense, whereas their normalizations may be compared reasonably.

Example 2.14 (Students’ deviation values). Suppose that a candidate got 75 points out of 100 in a screening test A. Obviously, the value 75 contains no information about his rank among the candidates. Suppose he got 62 points out of 100 in another screening test B. We know that comparison of two values 75 and 62 does not imply that he gets a higher rank in test A than in test B. In that case the normalized points are more informative. According to our practical experience the normalized point varies mostly between \(-3\) and \(3\). Since negative numbers are not convenient in bureaucracy and two-digit numbers are preferable, the deviation value is defined by
\[
y = 50 + 10 \bar{x} = 50 + 10 \frac{x - \bar{x}}{s_x}.
\]
As is easily seen, the mean and the standard deviation of the deviation values over all the candidates are 50 and 10, respectively. Thus, the deviation value varies mostly between 20 and 80. Moreover, being approximated by a normal distribution, the deviation value is useful for estimating the rank of a candidate and comparing the results of different tests. Historically, the students’ deviation value was introduced by a Japanese high school teacher as a reasonable scale of scholastic attainments.

Exercise 2.15. There were two screening tests A and B. The mean and the standard deviation of the points of all candidates of test A are 70 and 12, respectively. Likewise, those of test B are 50 and 8. A candidate got 75 points in test A and 62 in test B. Discuss the results by means of students’ deviation values.

2.5 Use of second or higher moments

Theorem 2.16 (Chebyshev inequality). Let \( x_1, x_2, \ldots, x_n \) be data of size \( n \), and let \( \bar{x} \) be the mean and \( s > 0 \) the standard variance. For \( k > 0 \) let \( N(k) \) be the number of data \( x_i \) satisfying \( |x_i - \bar{x}| \geq ks \). Then we have
\[
\frac{N(k)}{n} \leq \frac{1}{k^2}.
\]

Proof. Coming back to the definition (2.9), we divide the sum in the right-hand side into two parts as follows:
\[
x^2 = \frac{1}{n} \sum_{i: |x_i - \bar{x}| \geq ks} (x_i - \bar{x})^2 + \frac{1}{n} \sum_{i: |x_i - \bar{x}| < ks} (x_i - \bar{x})^2.
\]
In the first sum we have \((x_i - \bar{x})^2 \geq (ks)^2\) since \(x_i\) satisfies \(|x_i - \bar{x}| \geq ks\) and the second sum is always non-negative. Therefore \(s^2\) is estimated as

\[
s^2 \geq \frac{1}{n} \sum_{i|\xi_i - \bar{x}| \geq ks} (ks)^2 = \frac{N(k)}{n} (ks)^2.
\]
Dividing both sides by \((ks)^2\), which is non-zero by assumption, we obtain (2.17).

**Example 2.17.** It follows from the Chebyshev inequality that the number of data more than \(\pm 2s\) depart from the mean is less than \(1/2^2 = 1/4\) of the total number of data. Let us examine it by using the data of Team A. Recall that the mean and the standard deviation are given by

\[
\bar{x} = 179.8, \quad s = 5.37,
\]
respectively. Hence that the value \(x_i\) deviates from the mean more than \(\pm 2s\) means that \(x_i < 169.04\) or \(x_i > 190.54\).

There are 3 data satisfying this condition so that \(N(2) = 3\). Since the total number of data is \(n = 83\), the relative frequency of the data deviating from the mean more than \(\pm 2s\) is

\[
\frac{N(2)}{n} = \frac{3}{83} = 0.036.
\]
Indeed, the above relative frequency is less than \(1/4 = 0.25\) as is inferred from the Chebyshev inequality.

The Chebyshev inequality holds independently of the size of data and their shape of distribution, whereas it gives often a rather rough estimate. In fact, in the above example, the real frequency 0.036 is much smaller than 1/4 that follows from the Chebyshev inequality. Thus, the equality of (2.17) holds in very special cases.

Having introduced so far basic statistics such as the mean, variance, standard deviation, minimum, maximum, median, and so forth, we add a few more statistics. Given data \(x_1, x_2, \ldots, x_n\) let \(\bar{x}\) denote the mean as usual. For a natural number \(k\) the central moment of degree \(k\) is defined by

\[
m_k = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^k.
\]
The variance \(s^2\) is nothing else but the second central moment \(m_2\). Recall that the positive square root of it is the standard deviation \(s = \sqrt{m_2}\).

Using the cubic central moment \(m_3\) we define the skewness by

\[
\sqrt{\beta_1} = \frac{m_3}{s^3}.
\]
The somehow confusing symbol \(\sqrt{\beta_1}\) is common for statistics. We note that \(\sqrt{\beta_1}\) may take a negative value. In fact, skewness measures asymmetry of the distribution of data with respect to the mean. If the distribution has a heavier tail on the right, the skewness becomes a larger positive value. If the distribution has a heavier tail on the left, the skewness \(\sqrt{\beta_1}\) becomes a larger negative value. While, if the distribution is symmetric, we have \(\sqrt{\beta_1} = 0\).

Using the central moment of fourth order \(m_4\) we define the kurtosis by

\[
\beta_2 = \frac{m_4}{s^4}.
\]
This measures the grade of concentration of the data around the mean.

**Example 2.18.** The skewness and kurtosis of height of players in Team A are given by

\[
\sqrt{\beta_1} = 0.459, \quad \beta_2 = 3.038,
\]
respectively. The positive skewness suggests that the distribution of heights has a heavier tail on the right-side of the mean. The kurtosis near to 3 suggests that the distribution is similar to a normal distribution, see Example 2.19 below.

**Example 2.19.** In statistics the normal distribution is of fundamental importance. It is a continuous distribution given by the density function:

\[
f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right),
\]
where \(\mu\) is the mean and \(\sigma^2\) the variance, and is denoted by \(N(\mu, \sigma^2)\). The outline of \(f(x)\) is illustrated in Fig. 2.7. The skewness and kurtosis of the normal distribution \(N(\mu, \sigma^2)\) are independent of \(\mu\) and \(\sigma^2\), and are given by

\[
\sqrt{\beta_1} = 0, \quad \beta_2 = 3,
\]
respectively. These values are to be compared with the ones in Example 2.18, see also Fig. 2.2.
Remark 2.20. In some literatures, \( m_4/s^4 - 3 \) is taken to be the definition of kurtosis. This alternative definition is useful for checking similarity to the normal distribution, of which the kurtosis is 3 as shown in Example 2.19.

Remark 2.21. The normal distribution is observed often in real world. Suppose that a population obeys a normal distribution with mean \( \mu \) and standard deviation \( \sigma \), see Fig. 2.2. Then we have

(i) About 68% of the values lie within 1 standard deviation of the mean. In statistical notation, this is represented as \( \mu \pm 1\sigma \).

(ii) About 95% of the values lie within 2 standard deviations of the mean, that is, \( \mu \pm 2\sigma \).

(iii) About 99.7% of the values lie within 3 standard deviations of the mean, that is, \( \mu \pm 3\sigma \).

The above three facts are often called the *empirical rule* or the *68-95-99.7 rule*.

3. Description of Multi-Variate Data

3.1 Two-variate data and scatter plot

Let us start with two-variate data given by an \( n \times 2 \) data matrix:

\[
\begin{pmatrix}
  x_1 & y_1 \\
  x_2 & y_2 \\
  \vdots & \vdots \\
  x_n & y_n \\
\end{pmatrix}
\]

where \( x \) and \( y \) stand for the variables corresponding to measurement, and \( n \) is the size of data (the number of surveyed individuals). A pair of values \((x_i, y_i)\) is identified with a point in the \( xy \)-coordinate plane. Then, the data matrix (3.1) is transferred into a set of \( n \) points plotted in the coordinate plane, which is called the *scatter plot* or *scatter diagram* of the data.

A scatter plot is useful to check relationship between two variables. In this context, the relationship is called *correlation* in general. If the scatter plot is approximately along a straight line, the relationship is called *linear correlation*. We will discuss only linear correlation.

(i) If the scatter plot shows an uphill pattern from left to right, we say that the two variables are *positively correlated*. Namely, as the \( x \)-values increase (move right), the \( y \)-values increase (move up).

(ii) If the scatter plot shows a downhill pattern from left to right, two variables are *negatively correlated*. Namely, as the \( x \)-values increase (move right), the \( y \)-values decrease (move down).
Here is an example. The left diagram of Fig. 3.2 shows the scatter plot of height (horizontal axis) and weight (vertical axis) of players of Team A. Therein we may observe a trend of growing, which means that higher players are generally heavier. Of course, this trend is understood from the common-sense perspective. Likewise, the right diagram of Fig. 3.2 shows the scatter plot of age (horizontal axis) and height (vertical axis), where it seems difficult to find a growing or declining trend.

A scatter plot is useful to roughly grasp correlation but decision by looking easily leads to a mistake. A better treatment is to use the normalized data. Figure 3.3 shows the scatter plots of the normalized data of the original ones used in Fig. 3.2. Since the mean of the normalized data is 0, the scatter plot becomes a set of points distributed around the origin $\left(0, 0\right)$. Moreover, since the variance of normalized data is 1, the variability of points along the horizontal and vertical axes are unified.

**3.2 Correlation coefficient**

Two variables are correlated more strongly if points of the scatter plot are more tightly concentrated along a straight line. For proper judgement of the strength of correlation we need a statistic called the correlation coefficient.

Let $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ be two-variate data. The mean and variance of the variable $x$ are given by

$$
\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i, \quad s_x^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2.
$$

Similarly, for the variable $y$ we have

$$
\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i, \quad s_y^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2.
$$

We need a new statistic depending on both variables. The covariance of $x$ and $y$ is defined by

$$
s_{xy} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}).
$$
By definition we have

\[ s_{xy} = s_{yx}, \quad s_{xx} = s_x^2, \quad s_{yy} = s_y^2. \]

Expanding the right-hand side of (3.4), we obtain

\[
s_{xy} = \frac{1}{n} \sum_{i=1}^{n} (x_i y_i - \bar{x} \cdot \bar{y}) = \frac{1}{n} \sum_{i=1}^{n} x_i y_i - \frac{1}{n} \sum_{i=1}^{n} x_i \cdot \frac{1}{n} \sum_{i=1}^{n} y_i + \frac{1}{n} \sum_{i=1}^{n} \bar{x} \cdot \bar{y} = \frac{1}{n} \sum_{i=1}^{n} x_i y_i - \bar{x} \cdot \bar{y} = \frac{1}{n} \sum_{i=1}^{n} x_i y_i - \bar{x} \cdot \bar{y}.
\]

The sum in the last expression is the mean of the variable \( xy \), which is naturally denoted by \( \bar{xy} \). We thus come to the useful formula:

\[ s_{xy} = \bar{xy} - \bar{x} \cdot \bar{y}. \quad (3.5) \]

We say that \( x \) and \( y \) are positively correlated if the covariance is positive \( s_{xy} > 0 \). Similarly, \( x \) and \( y \) are negatively correlated if the covariance is negative \( s_{xy} < 0 \). Finally, \( x \) and \( y \) are uncorrelated if the covariance vanishes \( s_{xy} = 0 \).

We see from (3.4) that the covariance \( s_{xy} \) is more likely positive if there are a larger number of data \( (x_i, y_i) \) with \( (x_i - \bar{x})(y_i - \bar{y}) > 0 \) than those with \( (x_i - \bar{x})(y_i - \bar{y}) < 0 \). In other words, \( s_{xy} \) is more likely positive if more points are scattered in the upper right or lower left regions with respect to the mean point \( (\bar{x}, \bar{y}) \), see Fig. 3.4. In that case a growing trend of the scatter plot is more likely observed. A declining trend is similarly understood.

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{fig3.4.png}
\caption{Graphical understanding of the covariance.}
\end{figure}

In order to judge strength of the correlation we take normalized data. Let \( \bar{x} \) and \( \bar{y} \) be the normalized variables \( x \) and \( y \), respectively. The normalized data are given by

\[ \bar{x}_i = \frac{x_i - \bar{x}}{s_x}, \quad \bar{y}_i = \frac{y_i - \bar{y}}{s_y}. \quad (3.6) \]

Recall that the means of the normalized data are zero: \( \bar{x} = \bar{y} = 0 \). Applying the definition of covariance (3.4) to the pair of normalized variables \( (\bar{x}_i, \bar{y}_i) \), we obtain

\[
s_{\bar{x}\bar{y}} = \frac{1}{n} \sum_{i=1}^{n} (\bar{x}_i - \bar{x})(\bar{y}_i - \bar{y}) = \frac{1}{n} \sum_{i=1}^{n} \bar{x}_i \bar{y}_i = \frac{1}{n} \sum_{i=1}^{n} x_i - \bar{x} \frac{y_i - \bar{y}}{s_y} = \frac{1}{n} \sum_{i=1}^{n} x_i - \bar{x} \frac{y_i - \bar{y}}{s_y}.
\]

The last term being written in terms of the covariance of \( x \) and \( y \), we come to an important formula:

\[ s_{\bar{x}\bar{y}} = \frac{s_{xy}}{s_x s_y}. \quad (3.7) \]

The above statistic is called the correlation coefficient of \( (x, y) \) and is denoted by \( r = r_{xy} \). In other words, the correlation coefficient is defined by

\[ s_{\bar{x}\bar{y}} = \frac{s_{xy}}{s_x s_y}. \]
\[
\begin{align*}
    r &= r_{xy} = \frac{s_{xy}}{s_x s_y}, \\
    \text{Proof.} 
\end{align*}
\] (3.8)

In short, the correlation coefficient is the normalized covariance.

Of course, the signature of the correlation coefficient coincides with the one of covariance. We say that a pair of variables \((x, y)\) is **positively correlated** if they have a positive correlation. Similarly, \((x, y)\) is **negatively correlated** if they have a negative correlation. If the correlation coefficient is zero, there is no correlation between the two variables.

**Theorem 3.1.** For the correlation coefficient of two variables \(x, y\) we have

\[
-1 \leq r_{xy} \leq 1. 
\] (3.9)

**Proof.** From the definition (3.8) we see immediately that \(r_{xy} = r_{yx}\). For the inequality in (3.9) it is sufficient to show that \(r_{xy}^2 \leq 1\). As usual, the normalized variables of \(x\) and \(y\) are denoted by \(\tilde{x}\) and \(\tilde{y}\), respectively. We start with the obvious inequality:

\[
\sum_{i=1}^{n} (t\tilde{x}_i - \tilde{y}_i)^2 \geq 0, \quad t \in \mathbb{R}.
\]

Expanding the left-hand side, we have

\[
\left(\sum_{i=1}^{n} \tilde{x}_i^2\right)t^2 - 2\left(\sum_{i=1}^{n} \tilde{x}_i \tilde{y}_i\right)t + \left(\sum_{i=1}^{n} \tilde{y}_i^2\right) \geq 0.
\]

Dividing both sides by \(n\) and using \(\bar{x} = \bar{y} = 0\), we obtain

\[
s_x^2 t^2 - 2s_{xy} t + s_y^2 \geq 0.
\] (3.10)

Using \(s_x = s_y = 1\) and \(s_{xy} = r_{xy}\), we come to

\[
t^2 - 2r_{xy} t + 1 \geq 0,
\]

which holds for all real numbers \(t \in \mathbb{R}\). Hence the discriminant \(D = (-2r_{xy})^2 - 4 \leq 0\), from which \(r_{xy}^2 \leq 1\) follows.

As is stated in Theorem 3.1, the correlation coefficient \(r\) always lies between \(-1\) and \(+1\). We can interpret various values of \(r\) as follows:

(i) A correlation \(r\) exactly equal to \(-1\) indicates a perfect negative (linear) correlation (Exercise 3.7).
(ii) A correlation \(r\) close to \(-1\) indicates a strong negative correlation.
(iii) A correlation \(r\) close to \(0\) means no linear correlation.
(iv) A correlation \(r\) close to \(+1\) indicates a strong positive correlation.
(v) A correlation \(r\) exactly equal to \(+1\) indicates a perfect positive (linear) correlation (Exercise 3.7).

Most statisticians accept that the correlation is strong if the correlation coefficient is above \(+0.60\) or below \(-0.60\). Note however that the correlation coefficient is applied only for linear correlation, see Remark 3.3 below.

**Example 3.2.** Table 3.1 shows correlation coefficients of height, weight and age of players in Team A. The correlation coefficient 0.628 is not very strong but is enough to indicate the trend of growing along a straight line, see the left diagram in Fig. 3.3. The correlation coefficient of age and height is almost zero, as is suggested by the scatter plot, see the right diagram in Fig. 3.3.

| Covariance       | Correlation coefficient |
|------------------|-------------------------|
| height and weight | 28.27                   | 0.628                   |
| age and height    | -3.46                   | -0.130                  |
| age and weight    | -1.33                   | -0.032                  |

**Remark 3.3.** Even when the correlation coefficient is almost zero, we can not infer that there is no correlation between the two variables. Figure 3.5 shows two scatter plots of which the correlation coefficients are 0.043 (left) and 0.082 (right), whereas both scatter plots suggest correlations. In the left case, the data are scattered along an ellipse suggesting a “quadratic” relation between two variables. The correlation coefficient reflects only “linear” correlation so it is useless for non-linear relations. In the right case, we see that most data are scattered clearly along a straight line but there are a few extremal data. In fact, the correlation coefficient of the data except the extremal ones is 0.915 indicating a very strong linear correlation. It is noted that the correlation coefficient is sensitive to extremal data.
Remark 3.4. The correlation is a unitless measure. This means that if we change the units of \( x \) or \( y \), the correlation does not change. For example, changing the height (\( y \)) from centimeters to inches will not affect the correlation between the age and height. Also, as explained in Theorem 3.1, the correlation does not change after switching the variables \( x \) and \( y \) in the data set.

Remark 3.5. Clearly, condition that \( s_x > 0 \) and \( s_y > 0 \) is necessary to define the correlation coefficient. If \( s_x = 0 \) or \( s_y = 0 \), then data corresponding to the variable \( x \) or \( y \) are constant. In that case our original question of finding a growing or declining trend of a scatter plot does not make sense.

Exercise 3.6. For two variables \( x, y \) show that \( |s_{xy}| \leq s_x s_y \).

Exercise 3.7. Let \((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\) be two-variate data of size \( n \) and assume that \( s_x > 0 \) and \( s_y > 0 \). Show that all points in the scatter plot lie on a straight line with positive slope if and only if the correlation coefficient \( r_{xy} = 1 \). Similarly, show that the scatter plots lie on a straight line with negative slope if and only if \( r_{xy} = -1 \).

Exercise 3.8. Consider two variables \((x, y)\) and let \( r_{xy} \) be the correlation coefficient. For constant numbers \( a \) and \( b \) with \( a \neq 0 \) set \( x' = ax + b \). Show that

\[
r_{x'y} = \begin{cases} 
  r_{xy}, & \text{if } a > 0, \\
  -r_{xy}, & \text{if } a < 0.
\end{cases}
\]

3.3 Regression analysis

There are many problems transformed into an input-output model. Let us consider a system which receives an input and yields an output, where the system is often a black box with no detailed information about its operation. Here we consider a \( p \)-dimensional vector \((x_1, \ldots, x_p)\) as an input and a single variable \( y \) as an output.

\[
\begin{align*}
\text{Input} & \quad \rightarrow \quad \text{System} & \quad \rightarrow \quad \text{Output} \\
(x_1, \ldots, x_p) & \quad \rightarrow \quad (\text{Black Box}) & \quad \rightarrow \quad y \\
\end{align*}
\]

Mathematically a system is an unknown function:

\[
y = f(x_1, x_2, \ldots, x_p).
\]

Given a set of input-output data, we look for a function \( y = f(x_1, x_2, \ldots, x_p) \) which explains the data. For example, a manager of a beer company could make a good plan if the sales of beer \( y \) could be expected in terms of advertising cost \( x_1 \) and temperature \( x_2 \). But we naturally wonder there is no fundamental principle for this problem. Our approach consists of collecting data of three variables \((x_1, x_2, y)\) and looking for a function \( y = f(x_1, x_2) \) which recovers the data with reasonable accuracy.

To formulate our problems we need some notions and notations. A variable that we wish to predict is called a target variable or independent variable. While a variable that are used for calculating the target variable is called an explanatory variable, controlled variable or dependent variable. Let \( y \) be a target variable and \( x_1, x_2, \ldots, x_p \) be a set of explanatory variables. Given \((p + 1)\)-variate data, which is usually given in the form of of \( n \times (p + 1) \) data matrix:
our problem is to find a function \( y = f(x_1, x_2, \ldots, x_p) \) which recovers the given data. However, in general we can not hope to get a function \( y = f(x_1, x_2, \ldots, x_p) \) which reproduces all the data exactly. First of all, it is impossible if there are two data showing that the system (3.11) yields different outputs from the same inputs. In that case we might hope to add more explanatory variables to determine the function, but this strategy is not so realistic and hopeful because additional variables are often uncontrollable and unmeasurable. In fact, in a practical experiment or observation we can not specify all the variables which might affect the output. It is therefore essential to find a function \( y = f(x_1, x_2, \ldots, x_p) \) which recovers the data with reasonable accuracy. In other words, we allow an error term \( \epsilon \) to justify the data in such a way that

\[
y = f(x_1, x_2, \ldots, x_p) + \epsilon.
\]

In this context the function \( y = f(x_1, x_2, \ldots, x_p) \) is called a regression model. In particular, \( y = f(x_1, x_2, \ldots, x_p) \) is called a linear regression model if it is a linear function:

\[
y = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \beta_0,
\]

where \( \beta_0, \beta_1, \ldots, \beta_p \) are real coefficients. On the other hand, a regression model is called a single regression model if there is just one explanatory variable, and a multiple regression model otherwise. In general, the methodology of constructing regression models is called regression analysis.

### 3.4 Regression lines and method of least squares

We consider a single linear regression model. Given two-variate data

\[
(x_1, y_1), \ldots, (x_i, y_i), \ldots, (x_n, y_n),
\]

our problem is to determine a linear function

\[
y = ax + b,
\]

which recovers the data with reasonable accuracy. Such a linear function is also called a regression line. Here we take \( x \) as the explanatory variable and \( y \) as the target variable. Accordingly, a value \( (x_i, y_i) \) in the data is understood in such a way that an input \( x = x_i \) yields an output \( y = ax_i + b \) by (3.15) but the observed value \( y_i \) appears with deviation or fluctuation caused by some uncontrolled effects. Define the deviation \( \epsilon_i \) by

\[
y_i = ax_i + b + \epsilon_i,
\]

see Fig. 3.6. We consider that the most reasonable model minimizes the total deviation. In fact, there are several ways of defining the total deviation. The sum of squared deviations:

\[
Q = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - ax_i - b)^2
\]

is the most fundamental for some theoretical and practical reasons. Thus, our task is to find the constants \( a \) and \( b \) that minimize \( Q = Q(a, b) \). This principle is called the method of least squares, tracing back to Gauss and Legendre. A
linear regression model or a regression line is usually obtained by means of the method of least squares.

We outline the argument of deriving the linear regression model. Since the sum of squared deviations \( Q = Q(a, b) \) is a quadratic function in \( a \) and \( b \) though it has a lengthy expression, the minimum is found by simple algebra of completing the square or by simple differential calculus. The essence is stated in the following

**Lemma 3.9.** Given \( n \) data \( (x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n) \) with \( s_x > 0 \), the quadratic function:

\[
Q(a, b) = \sum_{i=1}^{n} (y_i - ax_i - b)^2
\]

attains the minimum at \( a = a_0 \) and \( b = b_0 \) given by

\[
a_0 = \frac{s_{xy}}{s_x^2}, \quad b_0 = \bar{y} - a_0 \bar{x}.
\]

**Proof.** Expanding the right-hand side of (3.17), we obtain

\[
Q = \sum (y_i^2 + a^2x_i^2 + b^2 - 2ax_iy_i - 2by_i + 2abx_i)
\]

\[
= \sum y_i^2 + a^2 \sum x_i^2 + b^2n - 2a \sum x_iy_i - 2b \sum y_i + 2ab \sum x_i,
\]

where the sum is always taken over \( 1 \leq i \leq n \). Use of the mean, variance and covariance:

\[
\bar{x} = \frac{1}{n} \sum x_i, \quad \bar{x}_x = \frac{1}{n} \sum x_i^2 - \bar{x}^2,
\]

\[
\bar{y} = \frac{1}{n} \sum y_i, \quad \bar{y}_y = \frac{1}{n} \sum y_i^2 - \bar{y}^2,
\]

\[
s_{xy} = \frac{1}{n} \sum x_iy_i - \bar{x} \cdot \bar{y}
\]

is slightly helpful for a concise expression. In fact, after simple algebra we obtain

\[
Q = n(s_{xy}^2 + \bar{y}^2) + a^2n(s_{xx}^2 + \bar{x}^2) + b^2n - 2an(s_{xy} + \bar{x} \cdot \bar{y}) - 2bn\bar{y} + 2abn\bar{x}.
\]

We see from the form in (3.16) that \( Q = Q(a, b) \) takes a minimum. Then we need only to find the stationary points of \( Q(a, b) \). The partial derivatives are easily obtained as

\[
\frac{\partial Q}{\partial a} = 2an(s_{xy} + \bar{x}^2) - 2n(s_{xy} + \bar{x} \cdot \bar{y}) + 2bn\bar{x},
\]

\[
\frac{\partial Q}{\partial b} = 2bn - 2n\bar{y} + 2an\bar{x}.
\]

Thus, our task is to solve the linear system:

\[
\frac{\partial Q}{\partial a} = \frac{\partial Q}{\partial b} = 0.
\]

Indeed, \((a_0, b_0)\) given in (3.18) is a unique solution, which means that \( Q = Q(a, b) \) attains the minimum only thereat. \( \Box \)

**Remark 3.10.** In Lemma 3.9 we assume that \( s_x > 0 \). That \( s_x = 0 \) is equivalent to that \( x_1, x_2, \ldots, x_n \) are constant. If two-variate data \((x_i, y_i)\) has this property, the question of finding a regression model makes no sense.

**Remark 3.11.** An alternative proof of Lemma 3.9 is by an elementary algebra. Using the identity:

\[
\frac{1}{n} Q(a_0, b_0) = \bar{x}_x - \frac{s_{xy}}{s_x^2},
\]

we easily obtain

\[
\frac{1}{n} (Q(a, b) - Q(a_0, b_0)) = (a\bar{x} + b - \bar{y})^2 + \left( a\bar{x} - \frac{s_{xy}}{s_x} \right)^2 \geq 0.
\]

Thus we see that \( Q(a, b) \geq Q(a_0, b_0) \) holds for all \( a, b \).

**Theorem 3.12.** For two-variate data \((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\) the regression line is given by

\[
\frac{y - \bar{y}}{s_y} = \frac{x - \bar{x}}{s_x},
\]

(3.20)
where \( x \) is the explanatory variable and \( y \) the target variable. Similarly, the regression line with explanatory variable \( y \) and target variable \( x \) is given by

\[
\frac{x - \bar{x}}{s_x} = r_{yx} \frac{y - \bar{y}}{s_y}.
\]  

(3.21)

**Proof.** By definition, the regression line with explanatory variable \( x \) and target variable \( y \) is given by \( y = a_0 x + b_0 \), where \( a_0 \) and \( b_0 \) are given as in (3.18). Using the explicit expression in Lemma 3.9, we come to

\[
y - \bar{y} = \frac{s_y}{s_x} r_{xy} (x - \bar{x}).
\]

Furthermore, using the correlation coefficient \( r_{xy} = s_{xy}/(s_x s_y) \), we have

\[
y - \bar{y} = \frac{s_y}{s_x} r_{xy} (x - \bar{x}).
\]

Dividing both sides by \( s_y \), we obtain (3.20). The second half of the statement is obvious by exchanging the roles of \( x \) and \( y \), together with an obvious relation \( r_{yx} = r_{xy} \).

It is noted that two regression lines (3.20) and (3.21) pass the common point \((\bar{x}, \bar{y})\) but their slopes are different. In fact, for the ratio of their slopes we have

\[
\frac{s_x r_{xy}}{s_x r_{xy} s_x} \leq \frac{s_x}{s_x} r_{xy} \leq 1.
\]

In other words, the roles of explanatory and target variables are not symmetric.

**Example 3.13.** We examined the height and weight of players of Team A in previous subsections. Let us remind some statistics:

\[
\bar{x} = 179.8, \quad \bar{y} = 82.9, \quad s_x = 5.37, \quad s_y = 8.39, \quad r_{xy} = 0.628.
\]

Then the regression line with height as explanatory variable \( x \) and weight as target variable \( y \) is given by

\[
y = \frac{82.9}{8.39} x - \frac{179.8}{5.37},
\]

that is,

\[
y = 0.98x - 93.3.
\]  

(3.22)

Similarly, the regression line with weight as explanatory variable \( y \) and height as target variable \( x \) is given by

\[
x = \frac{179.8}{5.37} y - \frac{82.9}{8.39},
\]

that is,

\[
x = 0.40y + 146.6.
\]  

(3.23)

The above equation is equivalent to \( y = 2.5x - 366.5 \), of which the slope in the \( xy \)-coordinate plane is 2.5 and is bigger than the slope of (3.22) certainly, see Fig. 3.7.
In Example 3.13, getting a strong linear correlation, we found the regression lines (3.22) and (3.23). Both are the best-fitting lines to data. The regression line (3.22) is used to predict a $y$-value from a given $x$-value. In other words, using $x$-variable whose data are easily observed or collected, we can predict the value of $y$-variable that is difficult or impossible to measure. The regression line (3.23) is used when the roles of $x$- and $y$-variables are exchanged. This idea works well as long as $x$ and $y$ are correlated.

Finally, we mention an important remark on application of regression lines. We know that the regression line is determined by means of five statistics $\bar{x}, \bar{y}, s_x, s_y, s_{xy}$. Hence, without looking at a scatter plot we can get the regression line as a result of simple calculus. Figure 3.8 shows two examples of scatter plots of normalized data and their regression lines with explanatory variable $x$ (horizontal axis) and target variable $y$ (vertical axis). The correlation coefficients are 0.756 (left) and 0.415 (right), both of which show positive correlation. Looking at the left scatter plot we are easily convinced that application of regression line is improper because the scattered point obeys more likely a quadratic curve. On the other hand, the right scatter plot shows that most data are scattered along a straight line which is different from the regression line. This is caused by a few extremal data. In this case we must examine the extremal data carefully. To avoid a risk of misuse of a regression line it is recommended to look at a scatter plot.

**Exercise 3.14.** For two-variate data $(x, y)$ let $L_1$ be the regression line with explanatory variable $x$ and target variable $y$, and $L_2$ the one with explanatory variable $y$ and target variable $x$. Show that the modulus of the slope of $L_2$ is greater than or equal to that of $L_1$.

**Exercise 3.15.** Let $\theta (0 \leq \theta \leq \pi/2)$ be the intersection angle of the regression line with explanatory variable $x$ and target variable $y$, and the one with explanatory variable $y$ and target variable $x$. Show that $\tan \theta = \frac{r_{xy}}{r_{yx}} \frac{s_x s_y}{s_x^2 + s_y^2}$.

Then prove that the intersection angle becomes closer to $\pi/2$ as the correlation between $x$ and $y$ becomes weaker, and that it becomes closer to 0 as the correlation between $x$ and $y$ becomes stronger.

### 3.5 Description of general multi-variate data

Now we discuss general $p$-variate data of variables $x_1, \ldots, x_j, \ldots, x_p$. We start with a data matrix of the form:

$$X = \begin{bmatrix}
  x_{11} & \cdots & x_{1j} & \cdots & x_{1p} \\
  \vdots & \ddots & \vdots & \ddots & \vdots \\
  x_{i1} & \cdots & x_{ij} & \cdots & x_{ip} \\
  \vdots & \ddots & \vdots & \ddots & \vdots \\
  x_{n1} & \cdots & x_{nj} & \cdots & x_{np}
\end{bmatrix}.$$  

(3.24)

The $i$th row of $X$ gives rise to a $p$-dimensional vector denoted by

$$x_i = [x_{i1}, \ldots, x_{ij}, \ldots, x_{ip}].$$  

(3.25)

As usual we identify $x_i$ with a point in $p$-dimensional coordinate space. Our main interest lies in how those points corresponding to the data (3.24) are distributed in the $p$-dimensional coordinate space. In the previous subsections we studied the case of two-variate data ($p = 2$), where visualization by a scatter plot is useful. In case of general $p$-variate data, direct observation of the scatter plot is not easy and reduction of dimension becomes important.
Suppose we are given \( p \)-variate data in the form of data matrix as in (3.24). Focusing on a variable \( x_j \), chosen from the \( p \) variables, we obtain a single-variate data:

\[
x_{ij}, \ldots, x_{ij}, \ldots, x_{nj},
\]

which appear as the \( j \)th column of the data matrix \( X \). Then the mean and variance of the variable \( x_j \) are defined by

\[
\bar{x}_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij},
\]

\[
\sigma^2_{x_j} = \frac{1}{n} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2,
\]

respectively. Similarly, for two variables \( x_j \) and \( x_k \) we define their covariance and correlation coefficient by

\[
s_{x_j,x_k} = \frac{1}{n} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k),
\]

\[
r_{x_j,x_k} = \frac{s_{x_j,x_k}}{\sigma_{x_j} \sigma_{x_k}},
\]

respectively. From now on, avoiding annoying symbols we write

\[
s_j^2 = s_{x_j}^2, \quad s_{jk} = s_{x_j,x_k}, \quad r_{jk} = r_{x_j,x_k}.
\]

We note by definition that

\[
s_{jj} = s_{x_j}^2, \quad r_{jj} = 1.
\]

With these statistics we define two \( p \times p \) matrices by

\[
\Sigma = \begin{bmatrix}
s_{11} & \cdots & s_{1p} \\
\vdots & \ddots & \vdots \\
s_{p1} & \cdots & s_{pp}
\end{bmatrix},
\]

\[
R = \begin{bmatrix}
r_{11} & \cdots & r_{1p} \\
\vdots & \ddots & \vdots \\
r_{p1} & \cdots & r_{pp}
\end{bmatrix}.
\]

The former is called the variance-covariance matrix and the latter the correlation matrix. Both matrices are symmetric in the sense that they are invariant under transposition, namely, \( s_{jk} = s_{kj} \) and \( r_{jk} = r_{kj} \). Note also that the diagonal entries of the correlation matrix are all \( r_{jj} = 1 \). The variance-covariance matrix and correlation matrix are fundamental in multi-variate analysis.

It is noticeable that the variance-covariance matrix and correlation matrix are derived directly from the data matrix \( X \) by means of matrix operations. Let \( J \) be the \( n \times n \) matrix whose entries are all one, i.e.,

\[
J = \begin{bmatrix}
1 & 1 & \cdots & 1 \\
1 & 1 & \cdots & 1 \\
\vdots & \vdots & \ddots & \vdots \\
1 & 1 & \cdots & 1
\end{bmatrix}.
\]

Calculating \( JX \) and comparing the mean (3.26), we obtain

\[
\frac{1}{n} (JX)_{ij} = \frac{1}{n} \sum_{k=1}^{n} (J)_{ik}(X)_{kj} = \frac{1}{n} \sum_{k=1}^{n} x_{kj} = \bar{x}_j.
\]

We set

\[
Y = X - \frac{1}{n} JX.
\]

Then we have

\[
(Y)_{ij} = \left( X - \frac{1}{n} JX \right)_{ij} = x_{ij} - \bar{x}_j.
\]

Since \( Y \) is an \( n \times p \) matrix, the product \( Y^T Y \) is defined and becomes a \( p \times p \) matrix. The \((j,k)\) entry of \( Y^T Y \) is given by

\[
(Y^T Y)_{jk} = \sum_{i=1}^{n} (Y^T)_{ji}Y_{ik} = \sum_{i=1}^{n} (Y)_{ij}Y_{ik},
\]

and, with the help of (3.32) we obtain
\[(Y^T Y)_{jk} = \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k).\]

In view of the covariance of two variables \(x_j\) and \(x_k\) given in (3.28), we see that

\[s_{jk} = \frac{1}{n} (Y^T Y)_{jk}.\]

Consequently, the variance-covariance matrix \(\Sigma\) in (3.30) becomes

\[
\Sigma = \frac{1}{n} Y^T Y = \frac{1}{n} \left( X - \frac{1}{n} J X \right)^T \left( X - \frac{1}{n} J X \right). \tag{3.34}
\]

For the correlation matrix we prepare a \(p \times p\) matrix defined by

\[
D = \begin{bmatrix}
\sqrt{s_{11}} & 0 & \cdots & 0 \\
0 & \sqrt{s_{22}} & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & \sqrt{s_{pp}}
\end{bmatrix}, \tag{3.35}
\]

where the diagonal entries are the standard deviations and the off-diagonal entries are all zero. Note that \(D^{-1}\) is the diagonal matrix with diagonal entries are all inverse of those of \(D\). Then

\[Z = YD^{-1}\]

becomes an \(n \times p\) matrix whose \((i, j)\) entry is given by

\[z_{ij} = \frac{1}{\sqrt{s_{jj} s_{kk}}} \sum_{k=1}^{p} (Y)_{ik}(D^{-1})_{jk} = (Y)_{ij}(D^{-1})_{jk} = \frac{x_{ij} - \bar{x}_j}{\sqrt{s_{jj}}}.\]

In other words, \(z_{ij}\) is the normalization of \(x_{ij}\) and the matrix \(Z\) itself is the normalization of the data matrix. Moreover, \(Z^T Z\) becomes a \(p \times p\) matrix whose \((j, k)\) entry is given by

\[(Z^T Z)_{jk} = \sum_{i=1}^{n} (Z)_j (Z)_k = \sum_{i=1}^{n} z_{ij} z_{ik} = \sum_{i=1}^{n} \frac{x_{ij} - \bar{x}_j}{\sqrt{s_{jj}}} \frac{x_{ik} - \bar{x}_k}{\sqrt{s_{kk}}} = r_{jk}.
\]

We then see from (3.28) and (3.29) that

\[
\frac{1}{n} (Z^T Z)_{jk} = \frac{1}{\sqrt{s_{jj} s_{kk}}} \frac{1}{n} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k) = \frac{s_{jk}}{\sqrt{s_{jj} s_{kk}}} = \frac{s_{jk}}{s_{jk}} = r_{jk},
\]

which is the correlation coefficient. Thus, from the definition (3.30) we obtain

\[R = \frac{1}{n} Z^T Z = \frac{1}{n} (YD^{-1})^T (YD^{-1}) = \frac{1}{n} D^{-1} Y^T Y D^{-1}.\]

Finally, in view of (3.31) we obtain the formula for the correlation matrix \(R\):

\[
R = \frac{1}{n} D^{-1} \left( X - \frac{1}{n} J X \right)^T \left( X - \frac{1}{n} J X \right) D^{-1}. \tag{3.36}
\]

Moreover, being combined with (3.34), we come to the basic identity linking the variance-covariance and correlation matrices:

\[R = D^{-1} \Sigma D^{-1}.\]

Summing up, we claim the following result.

**Theorem 3.16.** Let \(X\) be an \(n \times p\) data matrix as in (3.24). Then the variance-covariance matrix and the correlation matrix are respectively given by

\[
\Sigma = \frac{1}{n} \left( X - \frac{1}{n} J X \right)^T \left( X - \frac{1}{n} J X \right), \quad R = D^{-1} \Sigma D^{-1},
\]

where \(J\) is the all-one matrix and \(D\) is the diagonal matrix consisting of the standard deviations of variables as in (3.35).

### 3.6 Multi-variate regression analysis

Introducing the method of least squares, we derived in Sect. 3.4 the regression line from two-variate data \((x, y)\), where \(x\) and \(y\) are explanatory and target variables, respectively. In this subsection we deal with a general case of
(p + 1)-variate data \((x_1, x_2, \ldots, x_p, y)\), where \(x_1, x_2, \ldots, x_p\) are explanatory variables and \(y\) is a target variable.

We start with an \(n \times (p + 1)\) data matrix given by

\[
\begin{bmatrix}
  x_{11} & x_{12} & \cdots & x_{1p} & y_1 \\
  \vdots & \vdots & & \vdots & \vdots \\
  x_{n1} & x_{n2} & \cdots & x_{np} & y_n
\end{bmatrix}
\]  

Our goal is to derive a multi-linear regression model:

\[ y = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \beta_0 \]  

by means of the method of least squares. As before, the \(i\)th value \((x_{i1}, x_{i2}, \ldots, x_{ip}, y_i)\) is understood in such a way that an input \((x_{i1}, x_{i2}, \ldots, x_{ip})\) yields the output \(y = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \beta_0\) according to (3.38) but an observed value \(y_i\) is deviated from it by uncontrolled effects. The deviation \(\epsilon_i\) is defined by

\[ y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \beta_0 + \epsilon_i. \]  

Then we will minimize the sum of squared deviations:

\[ Q = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - (\beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \beta_0))^2. \]  

This is the principle of the method of least squares. In fact, as \(Q = Q(\beta_1, \ldots, \beta_p, \beta_0)\) is a quadratic function we may apply a similar argument as in the case of \(p = 1\). In order to overcome the difficulty caused by the number of variables we employ matrix notation.

We first note that in the right-hand side of (3.38) the roles of \(\beta_1, \beta_2, \ldots, \beta_p, \beta_0\) are not equal. It is then convenient to introduce a dummy variable \(x_0\). The data corresponding to this new variable is set to be all one. Let \(X\) be the data matrix associated with the variables \(x_0, x_1, \ldots, x_p\) and \(y\) the one associated to \(y\). In fact, \(X\) becomes an \(n \times (p + 1)\) matrix and \(y\) an \(n \times 1\) matrix or an \(n\)-dimensional column vector:

\[
X = \begin{bmatrix}
  x_{10} & x_{11} & \cdots & x_{1p} \\
  \vdots & \vdots & & \vdots \\
  x_{n0} & x_{n1} & \cdots & x_{np}
\end{bmatrix}, \quad
y = \begin{bmatrix}
  y_1 \\
  \vdots \\
  y_n
\end{bmatrix}
\]  

where \(x_{i0} = 1\) for all \(i\). Next we define \((p + 1)\)-dimensional column vector by

\[
\beta = \begin{bmatrix}
  \beta_0 \\
  \beta_1 \\
  \vdots \\
  \beta_p
\end{bmatrix}
\]

Our problem is to determine \(\beta\) from \(X\) and \(y\). With the above matrix notations the deviation (3.39) becomes

\[ \epsilon_i = (y - X\beta)_i, \]  

where the right-hand side is the \(i\)th entry of \(n\)-dimensional vector \(y - X\beta\). Then the sum of squared deviations is given by the norm and inner product:

\[ Q = \sum_{i=1}^{n} \epsilon_i^2 = \| y - X\beta \|^2 = (y - X\beta, y - X\beta). \]  

The above simple expression helps our argument very much. Expanding the right-hand side, we obtain

\[ Q = (y, y) - 2(y, X\beta) + (X\beta, X\beta) = (y, y) - 2(X^T y, \beta) + (X^T X\beta, \beta). \]

It follows from the general theory that \(Q = Q(\beta)\) attains the minimum at a stationary point. Stationary points are characterized by the linear system:
\[ \frac{\partial Q}{\partial \beta_0} = \frac{\partial Q}{\partial \beta_1} = \cdots = \frac{\partial Q}{\partial \beta_p} = 0. \]  
\[ (3.44) \]

On the other hand, the partial derivative of \( Q \) is easily computed. Let \( e_j \) be the \((p + 1)\)-dimensional vector whose \( j \)th entry is one and the others are all zero. Then we have

\[ \frac{\partial Q}{\partial \beta_j} = -2(X^T y, e_j) + 2(X^T X\beta, e_j) = (-2X^T y + 2X^T X\beta, e_j). \]

from which we see that the linear system (3.44) is equivalent to

\[ -2X^T y + 2X^T X\beta = 0, \]

or equivalently,

\[ X^T X\beta = X^T y. \]  
\[ (3.45) \]

The above equation is often called the normal equation. Assuming that the matrix \( X^T X \) has the inverse, we come to a unique solution to (3.44), that is,

\[ \beta_0 = (X^T X)^{-1} X^T y. \]  
\[ (3.46) \]

Consequently, \( \beta = \beta_0 \) is the unique point at which \( Q = Q(\beta) \) attains the minimum. Summing up the above argument we come to the following statement.

**Theorem 3.17.** Assume that \((p + 1)\)-variate data are given by a data matrix as in (3.37). Introduce a dummy variable \( x_0 \) and set the corresponding data to be all one. Let \( X \) and \( y \) be data matrices defined as in (3.41), and assume that \( X^T X \) has the inverse. Then the multilinear regression model (3.38) that minimizes the sum of squared deviations is given by (3.46).

**Remark 3.18.** During the above argument we cannot drop the condition that the matrix \( X^T X \) has the inverse. If the size of data is large, the matrix \( X^T X \) has the inverse most probably in practice. On the other hand, it is proved that \( X^T X \) has no inverse if \( p > n \). Thus, the case where the number of variables exceeds the size of data requires an advanced methodology.

It is instructive to check directly that \( Q = Q(\beta) \) attains the minimum at \( \beta = \beta_0 \) given by (3.46). We start with (3.43):

\[ Q(\beta) = \| y - X\beta \|^2 \]
\[ = \| y - X\beta_0 + X\beta - X\beta \|^2 \]
\[ = \| y - X\beta_0 \|^2 + \| X\beta_0 - X\beta \|^2 + 2\langle y - X\beta_0, X\beta - X\beta \rangle \]
\[ = \| y - X\beta_0 \|^2 + \| X\beta_0 - X\beta \|^2 + 2\langle X^T (y - X\beta_0), \beta - \beta \rangle. \]  
\[ (3.47) \]

Since \( \beta_0 \) fulfills (3.45), we have \( X^T (y - X\beta_0) = 0 \) and hence the inner product in the last expression of (3.48) vanishes. We then have

\[ Q(\beta) = \| y - X\beta_0 \|^2 + \| X\beta_0 - X\beta \|^2 = Q(\beta_0) + \| X\beta_0 - X\beta \|^2 \geq Q(\beta_0). \]

Apparently, the equality happens only when \( X\beta_0 = X\beta \). But \( X\beta_0 = X\beta \) does not imply \( \beta = \beta_0 \) in general. If \( X^T X \) is invertible, we may obtain \( \beta = \beta_0 \). In that case \( Q = Q(\beta) \) attains the minimum at \( \beta = \beta_0 \) and the minimum is attained only at \( \beta = \beta_0 \).

**Example 3.19.** In Sect. 3.3 we derived the linear regression model for two-variate data. Of course, the method described in this subsection covers the case of \( p = 1 \). It is instructive to apply the matrix method to the case of \( p = 1 \). We start with two-variate data

\[ (x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n), \]

where \( y \) is the target variable and \( x \) is the explanatory variable. Introduce a dummy variable \( x_0 \) and rewrite \( x \) by \( x_1 \). Then the data matrices \( X \) and \( y \) in (3.41) take the forms:

\[ X = \begin{bmatrix} x_{10} & x_{11} \\ x_{20} & x_{21} \\ \vdots & \vdots \\ x_{n0} & x_{n1} \end{bmatrix}, \quad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \]

respectively. Then by direct calculation we have

\[ X^T X = \begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix} = \begin{bmatrix} n & nx \ \\ nx & nx^2 + nx^3 \end{bmatrix}. \]

It is known that \( X^T X \) has the inverse if and only if \( x_i \neq 0 \). This is equivalent to that the data of the variable \( x \) are not constant. Under this condition we have

\[ \frac{\partial Q}{\partial \beta_0} = \frac{\partial Q}{\partial \beta_1} = \cdots = \frac{\partial Q}{\partial \beta_p} = 0. \]  
\[ (3.44) \]
\[
(X^T X)^{-1} = \frac{1}{n^2 s_x^4} \begin{bmatrix}
ns_x^2 + nx^2 & -nx \\
-nx & n
\end{bmatrix} = \frac{1}{n s_x^2} \begin{bmatrix} 
s_x^2 + x^2 & -x \\
-x & 1
\end{bmatrix}.
\]

On the other hand, since
\[
X^T y = \begin{bmatrix} 1 & \cdots & 1 \\
x_1 & \cdots & x_n \\
y_1 & \vdots & y_n
\end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n y_i \\
\sum_{i=1}^n x_i y_i \\
y_n
\end{bmatrix} = \begin{bmatrix} n \bar{y} \\
n s_{xy} + n \bar{x} \cdot \bar{y}
\end{bmatrix},
\]
we see from the formula (3.46) that
\[
\beta_0 = (X^T X)^{-1} X^T y = \frac{1}{n s_x^2} \begin{bmatrix} s_x^2 + x^2 & -x \\
-x & 1 \\
\bar{y} & n s_{xy} + n \bar{x} \cdot \bar{y}
\end{bmatrix} = \frac{1}{s_x^2} \begin{bmatrix} s_x^2 \bar{y} & -s_{xy} \bar{x} \\
\bar{y} & s_{xy}
\end{bmatrix}.
\]

Consequently, the desired linear regression model is given by
\[
y = \beta_1 x + \beta_0,
\]
where
\[
\beta_0 = \frac{1}{s_x^2} (s_x^2 \bar{y} - s_{xy} \bar{x}) = \bar{y} - \frac{s_{xy}}{s_x^2} \bar{x}, \quad \beta_1 = \frac{s_{xy}}{s_x^2}.
\]

Of course, the result coincides with the one stated in Theorem 3.12.

4. Foundations of Probability

4.1 Events and probability

An event is the result of an observation or experiment, for which we can clearly decide whether or not it occurs. When the occurrence is not predicted with total certainty, we are interested in how likely it occurs. A probability is a scale to measure the likelihood by means of a real number between 0 and 1.

To be slightly more precise, we need a sample point, that is an outcome of observation or experiment which are indivisible and primary. Collecting all sample points, we form a sample space often denoted by \( \Omega \). Then an event \( A \) is understood as a subset of \( \Omega \), namely, a set of sample points. The probability that an event \( A \) occurs is denoted by \( P(A) \).

Example 4.1 (Coin tossing). In coin tossing we observe two possibilities, heads or tails. By convention we use numbers 1 and 0 for heads and tails, respectively. Then the sample space of coin toss becomes
\[
\Omega = \{0, 1\}.
\]

Since there are four subsets of \( \Omega \), we have four events for coin tossing:
\[
\emptyset, \quad \{0\}, \quad \{1\}, \quad \Omega = \{0, 1\}.
\]

In general, \( \emptyset \) stands for an event containing no sample point and is called an empty event or null event. While, the sample space \( \Omega \) itself is an event called the whole event. Since \( \emptyset \) never occurs and \( \Omega \) occurs with total certainty, we have
\[
P(\emptyset) = 0, \quad P(\Omega) = 1.
\]

Assuming that the coin is fair, we understand by symmetry that the probabilities of heads and tails are equal. Hence we have
\[
P(\{0\}) = P(\{1\}) = \frac{1}{2}.
\]

Remark 4.2. An event consisting of a single sample point is called an elementary event. Strictly speaking, a sample point \( \omega \in \Omega \) and an elementary event \( \{\omega\} \subset \Omega \) are conceptually different as a probability is given to an elementary event but not to a sample point. Nevertheless, we occasionally write \( P(\{\omega\}) = P(\omega) \) for simple notation.

Example 4.3 (Dice rolling). For rolling a die (with six sides) we may set
\[
\Omega = \{1, 2, 3, 4, 5, 6\}.
\]

For example, rolling a 1 is an elementary event denoted by \( \{1\} \), and rolling an even value is an event denoted by \( \{2, 4, 6\} \). By symmetry we have

\[
\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}.
\]
\[ P(\{1\}) = P(\{2\}) = P(\{3\}) = P(\{4\}) = P(\{5\}) = P(\{6\}) = \frac{1}{6} \]  

(4.2)

and

\[ P(\{2, 4, 6\}) = \frac{1}{2}. \]

Apparently it is not convenient to list up the probabilities of \(2^6\) events for dice rolling. The simple formula

\[ P(A) = \frac{|A|}{6}, \]

where \(|A|\) stands for the number of sample points of \(A\), is much more essential.

We find a common idea in (4.1) and (4.2), where an elementary event is given an equal probability. In general, consider a sample space \(\Omega\) with finitely many sample points. If every elementary event occurs equally likely, the probability of an event \(A \subset \Omega\) is given by proportion:

\[ P(A) = \frac{|A|}{|\Omega|}, \]  

(4.3)

where \(|A|\) denotes the number of sample points in \(A\). The formula (4.3), tracing back to the very early stage of combinatorial probability, is our starting point of combinatorial probability.

The essence of (4.3) is that an event is represented as a set and the probability is given by the ratio of “sizes” of sets. In fact, in combinatorial probability the size of a set is given by the number of elements. We may employ another measure for the “size” of a set. Let \(\Omega\) be a Euclidean domain and consider a trial to choose a point of \(\Omega\) randomly in such a way that every point is chosen equally likely. Then the probability that a point is chosen from a subset \(A \subset \Omega\) is naturally defined by the same formula (4.3), where the “size” of a set is measured in terms of length, area or volume according to the dimension of the Euclidean space. This idea is sometimes referred to as geometric probability.

**Example 4.4.** Consider an interval \(\Omega = [a, b]\) in the real line \(\mathbb{R}\) with \(a < b\). Choose a point of \(\Omega\) randomly in such a way that every point of \(\Omega\) is chosen equally likely. Let \(A\) be the event that a point is chosen from the interval \([\alpha, \beta]\), where \(a \leq \alpha \leq \beta \leq b\), see Fig. 4.1. Then we have

\[ P(A) = \frac{|A|}{|\Omega|} = \frac{\beta - \alpha}{b - a}, \]

where \(|A|\) denotes the length of \(A\). Note that the probability that a particular point (e.g., the mid-point of the interval) is chosen is zero.

![Fig. 4.1. Choosing a point from an interval \([a, b]\).](image)

**Example 4.5.** Let \(\Omega\) be a disc of radius \(R > 0\) and choose a point of \(\Omega\) randomly in such a way that every point of \(\Omega\) is chosen equally likely. Let \(A\) be a subset of \(\Omega\), see Fig. 4.2. Then the event that a point is chosen from \(A\), denoted by the same symbol, is given by

\[ P(A) = \frac{|A|}{|\Omega|} = \frac{|A|}{\pi R^2}, \]

where \(|A|\) denotes the area of \(A\). Strictly speaking, a subset \(A\) can not be arbitrary but we need to restrict ourselves to a measurable set, see Remark 4.16.

![Fig. 4.2. Choosing a point from a disc with radius \(R > 0\).](image)
4.2 Probability spaces

Now we mention the mathematical formulation of probability due to Kolmogorov. We start with a non-empty set $\Omega$ as a sample space. An event is a subset of $\Omega$. However, we do not require that every subset of $\Omega$ is an event, but we impose some conditions on the collection of events instead.

We need some operations on events. The complement of an event $A$, denoted by $A^c$, is the event that $A$ does not occur. For two events $A$ and $B$, the event that at least one of $A$ or $B$ occurs is called the union and is denoted by $A \cup B$. The event that both $A$ and $B$ occur is called the intersection and is denoted by $A \cap B$. Two events $A$ and $B$ are called disjoint if $A \cap B = \emptyset$.

![Fig. 4.3. Complement $A^c$, union $A \cup B$ and intersection $A \cap B$ (from left to right).]

**Remark 4.6.** There are some variants of notation in literatures. The complement $A^c$ is also denoted by $\bar{A}$. The union $A \cup B$ may be called the sum. The intersection $A \cap B$ is also called the product and is denoted by $AB$.

**Definition 4.7 (σ-field).** A collection $\mathcal{F}$ of subsets of $\Omega$ is called a $\sigma$-field if the following conditions are satisfied:

(i) $\emptyset \in \mathcal{F}$ and $\Omega \in \mathcal{F}$;
(ii) If $A \in \mathcal{F}$, then the complement $A^c \in \mathcal{F}$;
(iii) If $A_1, A_2, \ldots \in \mathcal{F}$, then the union $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

A $\sigma$-field is closed under countable intersection too, namely, for $A_1, A_2, \ldots \in \mathcal{F}$ the intersection $\bigcap_{i=1}^{\infty} A_i$ belongs to $\mathcal{F}$. It is a consequence of the definition that a $\sigma$-field is also closed under finite union and intersection (Exercise 4.10). It is essential that the events form a $\sigma$-field.

**Definition 4.8.** Let $\Omega$ be a non-empty set and $\mathcal{F}$ a $\sigma$-field over $\Omega$. A probability is a map (or function) $P: \mathcal{F} \to \mathbb{R}$ satisfying the following properties:

(i) $0 \leq P(A) \leq 1$ for all $A \in \mathcal{F}$;
(ii) $P(\Omega) = 1$;
(iii) If $A_1, A_2, \ldots \in \mathcal{F}$ are mutually disjoint, i.e., $A_i \cap A_j = \emptyset$ for $i \neq j$, then

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i).$$

**Definition 4.9.** A probability space is a triple $(\Omega, \mathcal{F}, P)$, where $\Omega$ is a non-empty set, $\mathcal{F}$ is a $\sigma$-field over $\Omega$, and $P$ is a probability defined on $\mathcal{F}$.

Once we are given a probability space $(\Omega, \mathcal{F}, P)$, the set $\Omega$ is called a sample space and an element $\omega \in \Omega$ a sample point. A subset $A$ of $\Omega$ which belongs to $\mathcal{F}$ is called an event, thus $\mathcal{F}$ is the space of events. Finally, $P(A)$ means the probability that an event $A$ occurs.

An event $A$ is called almost sure if $P(A) = 1$. By definition the whole event $\Omega$ is almost sure, but there might be many other almost sure events. An event $A$ is called almost impossible if $P(A) = 0$. We know that the empty event $\emptyset$ is almost impossible, but there might be many other almost impossible events, see Examples 4.4 and 4.5.

**Exercise 4.10.** Let $(\Omega, \mathcal{F}, P)$ be a probability space. Show that if $A$ and $B$ are events, so are $A \cap B$ and $A \cup B$.

**Exercise 4.11.** Consider the experiment in Example 4.3 of rolling a die. Set $A = \{2, 4, 6\}$, $B = \{1, 3, 5\}$ and $C = \{1, 2, 3, 4\}$. Show that $\{\emptyset, \Omega, A, B\}$ is a $\sigma$-field but $\{\emptyset, \emptyset, A, B, C\}$ are not.

**Exercise 4.12.** Let $(\Omega, \mathcal{F}, P)$ be a probability space. Prove the following assertions, where $A$ and $B$ are events.

1. $P(\emptyset) = 0$.
2. If $A \cap B = \emptyset$, then $P(A \cup B) = P(A) + P(B)$.
3. $P(A^c) = 1 - P(A)$, where $A^c$ is the complement.
4. If $A \subseteq B$, then $P(A) \leq P(B)$.
5. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. 

- **Exercise 4.13.** Consider a sample space $\Omega$ and a collection $\mathcal{F}$ of subsets of $\Omega$. Prove the following properties:
   - (i) If $A \in \mathcal{F}$, then $\emptyset \subseteq A$ and $A \subseteq \Omega$.
   - (ii) If $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$.
   - (iii) If $A_1, A_2, \ldots \in \mathcal{F}$, then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.
   - (iv) Let $A \in \mathcal{F}$ and $B \subseteq A$. Show that $A \cap B \in \mathcal{F}$.

- **Exercise 4.14.** Let $\Omega$ be a sample space and let $\mathcal{F}$ be a $\sigma$-field over $\Omega$. Prove that $\mathcal{F}$ is a field of subsets of $\Omega$. That is, if $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$ and $A \cap B \in \mathcal{F}$ for all $B \subseteq A$.

- **Exercise 4.15.** Let $\Omega$ be a sample space and let $\mathcal{F}$ be a $\sigma$-field over $\Omega$. Prove that $\mathcal{F}$ is a field of subsets of $\Omega$. That is, if $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$ and $A \cap B \in \mathcal{F}$ for all $B \subseteq A$.
Exercise 4.13. Let \((\Omega, \mathcal{F}, P)\) be a probability space and let \(A\) be an almost impossible event. Prove that \(P(A \cup B) = P(B)\) for any event \(B\).

Exercise 4.14. An experiment consists of tossing two dice.

1. Determine the probability space \((\Omega, \mathcal{F}, P)\).
2. Find the event \(A\) that the sum of the dots on the dice equals 8 and the probability \(P(A)\).
3. Find the event \(B\) that the sum of the dots on the dice is greater than 10 and the probability \(P(B)\).
4. Find the event \(C\) that the sum of the dots on the dice is greater than 12 and the probability \(P(C)\).

Exercise 4.15. Let \(\Omega = \{\omega_1, \omega_2, \ldots\}\) be a countable set and \(\mathcal{F}\) the set of all subsets of \(\Omega\). Let \(p_1, p_2, \ldots\) be a sequence such that \(p_i \geq 0\) and \(\sum_{i=1}^{\infty} p_i = 1\). Define \(P : \mathcal{F} \to \mathbb{R}\) by

\[
P(A) = \sum_{\omega \in A} p_i = \sum_{i=1}^{\infty} p_i 1_A(\omega),
\]

where \(1_A(\omega)\) is the indicator function defined by

\[
1_A(\omega) = \begin{cases} 
1, & \omega \in A, \\
0, & \text{otherwise}.
\end{cases}
\]

Show that \((\Omega, \mathcal{F}, P)\) is a probability space and determine the almost sure events.

Remark 4.16. In combinatorial probability, \(\Omega\) is a finite set and every subset of \(\Omega\) is an event. In Exercise 4.15, \(\Omega\) being an infinite set, every subset of \(\Omega\) is still an event. The idea of \(\sigma\)-field becomes essential in Examples 4.4 and 4.5. In fact, it is known that there is a subset of the real line of which the length is not determined and likewise that there is a subset of the plane of which the area is not determined. Hence for defining the probability by length or area, we need to avoid those pathological sets and to restrict ourselves to measurable sets that admit the length or area. It is a basic consequence of measure theory that the length or the area is well defined for all Borel sets which form the minimum \(\sigma\)-field containing all open subsets, for more details see the standard textbooks.

The above concept of probability is quite abstract and needs to relate to experiments in the real world. In fact, traditionally the probability was discussed in the context of strictly controlled experiments that can be repeated under identical conditions as many times as we like. Suppose that such an experiment is repeated \(n\) times. If an event \(A\) occurs \(n(A)\) times, then the relative frequency of \(A\) is defined by

\[
p_n(A) = \frac{n(A)}{n}.
\]

Obviously \(p_n(A)\) is not uniquely determined by \(n\). If the limit

\[
P(A) = \lim_{n \to \infty} p_n(A) = \lim_{n \to \infty} \frac{n(A)}{n}
\]

exists, we accept that \(P(A)\) is the probability of \(A\). Note that the above limit may not exist, and in addition, there are many situations in which the concepts of repeatability may not be valid.

Following the definition of probability space, the relative frequency of \(A\) possesses the following properties:

(i) \(0 \leq p_n(A) \leq 1\), where \(p_n(A) = 0\) if \(A\) occurs in none of the \(n\) repeated trials and \(p_n(A) = 1\) if \(A\) occurs in all of the \(n\) repeated trials.

(ii) If \(A\) and \(B\) are mutually exclusive (or disjoint) events, then

\[
p_n(A \cup B) = p_n(A) + p_n(B)
\]

and

\[
P(A \cup B) = \lim_{n \to \infty} \frac{n(A \cup B)}{n} = \lim_{n \to \infty} \frac{n(A)}{n} + \lim_{n \to \infty} \frac{n(B)}{n} = P(A) + P(B).
\]

Starting with a probability space, the basic limit formula (4.4) is proved within a somehow sophisticated argument of probability theory. The result is known as the law of large numbers (LLN), for details see standard textbooks.

Upon closing this subsection, we mention the basic concept of independent events.

Definition 4.17. A family of events \(\{A_\lambda; \lambda \in \Lambda\}\) is called independent if any finitely many events \(A_{\lambda_1}, \ldots, A_{\lambda_k}\) chosen from the family satisfies

\[
P\left(\bigcap_{i=1}^{n} A_{\lambda_i}\right) = \prod_{i=1}^{n} P(A_{\lambda_i}).
\]

In particular, two events \(A\) and \(B\) are independent if
\[
P(A \cap B) = P(A)P(B).
\]

Thus three events \( A, B \) and \( C \) are independent if and only if
\[
\begin{align*}
P(A \cap B \cap C) &= P(A)P(B)P(C), \\
P(A \cap B) &= P(A)P(B), \\
P(B \cap C) &= P(B)P(C), \\
P(C \cap A) &= P(C)P(A).
\end{align*}
\]

**Exercise 4.18.** Show that if three events \( A, B \) and \( C \) are independent, then \( A \) and \( B \cup C \) are independent.

**Exercise 4.19.** Let \( A \) and \( B \) be two events that are disjoint and independent. Show that \( P(A) = 0 \) or \( P(B) = 0 \). Thus, two disjoint events are not independent except the trivial case.

**Exercise 4.20.** In the experiment of rolling two fair dice, let \( A \) be the event that the first die is odd, \( B \) the event that the second die is odd, and \( C \) the event that the sum is odd. Show that the three events \( A, B \) and \( C \) are pairwise independent, but \( A, B \) and \( C \) are not independent.

### 4.3 Conditional probability

Let \((\Omega, \mathcal{F}, P)\) be a probability space. For two events \( A, B \in \mathcal{F} \) the **conditional probability** of \( A \) relative to \( B \) or under condition \( B \) is defined by
\[
P(A \mid B) = \frac{P(A \cap B)}{P(B)},
\]
whenever \( P(B) > 0 \). The conditional probability \( P(A \mid B) \) is interpreted as the probability that an event \( A \) occurs provided that we already know the occurrence of another event \( B \). Similarly, if \( P(A) > 0 \),
\[
P(B \mid A) = \frac{P(A \cap B)}{P(A)}.
\]

From (4.5) and (4.6), we obtain immediately the following

**Theorem 4.21** (Multiplicative rule). Let \( A \) and \( B \) be two events. If \( P(B) > 0 \), we have
\[
P(A \cap B) = P(B)P(A \mid B).
\]

Similarly, if \( P(A) > 0 \), we have
\[
P(A \cap B) = P(A)P(B \mid A).
\]

In computing the joint probability of events the useful relation:
\[
P(A \cap B) = P(A \mid B)P(B) = P(B \mid A)P(A)
\]
will be applied without an explanation.

**Theorem 4.22.** Let \( A \) and \( B \) be two events with \( P(B) > 0 \). Then \( A \) and \( B \) are independent if and only if \( P(A \mid B) = P(A) \).

**Proof.** By definition \( A \) and \( B \) are independent if and only if \( P(A \cap B) = P(A)P(B) \). Then the assertion is straightforward from the definition (4.5). \( \square \)

Thus, if \( A \) is independent of \( B \), then the probability that \( A \) occurs is unchanged by information on whether or not \( B \) occurs.

**Exercise 4.23.** Show that if two events \( A \) and \( B \) are independent, so are \( A \) and \( B^c \).

**Exercise 4.24.** Show the following relations, where \( A, B \) and \( C \) are events.
\[
\begin{align*}
(1) & \quad P(A \mid A) = 1, \\
(2) & \quad P(A \cap B \mid C) = P(B \mid C)P(A \mid B \cap C), \\
(3) & \quad P(A \cap B \cap C) = P(A \mid B \cap C)P(B \mid C)P(C).
\end{align*}
\]

**Theorem 4.25** (Bayes’ formula). Let \((\Omega, \mathcal{F}, P)\) be a probability space. Let \( B_1, \ldots, B_n \) be mutually disjoint events with \( P(B_i) > 0 \) such that
\[
\Omega = \bigcup_{i=1}^{n} B_i,
\]

\[
\sum_{i=1}^{n} P(B_i) = 1.
\]
Then for any event \( A \) with \( P(A) > 0 \) we have

\[
P(B_j | A) = \frac{P(B_j)P(A | B_j)}{\sum_{i=1}^{n} P(B_i)P(A | B_i)}.
\]

(4.10)

Proof. We first note that

\[
P(B_j | A) = \frac{P(A \cap B_j)}{P(A)}
\]

(4.11)

which is by definition. It follows from (4.9) that

\[
A = \bigcup_{i=1}^{n} (A \cap B_i).
\]

Since \( A \cap B_1, \ldots, A \cap B_n \) are mutually disjoint, we have

\[
P(A) = \sum_{i=1}^{n} P(A \cap B_i).
\]

(4.12)

On the other hand, for two events \( A \) and \( B_i \) we have

\[
P(A \cap B_i) = P(B_i)P(A | B_i)
\]

by the multiplicative rule. Then the numerator of the right-hand side in (4.11) becomes

\[
P(A \cap B_j) = P(B_j)P(A | B_j)
\]

(4.13)

and the denominator becomes

\[
P(A) = \sum_{i=1}^{n} P(B_i)P(A | B_i).
\]

(4.14)

Then the formula (4.10) is obtained by inserting (4.13) and (4.14) into (4.11).

□

Example 4.26 (Diagnostic test). Consider an infectious disease in a certain population \( \Omega \). The population \( \Omega \) is divided into two parts, infected \( D \) or health \( D^c \). There is a test \( T \) for the disease and the result is positive or negative. Again the population \( \Omega \) is divided into two parts, positive test result \( T^+ \) or negative test result \( T^- \). We now consider \( \Omega \) as a sample space and \( D, D^c, T^+ \) and \( T^- \) as events. The conditional probabilities \( P(T^+ | D) \) and \( P(T^- | D^c) \) are called the sensitivity and specificity of the test, respectively. Those values may be obtained in a laboratory. The conditional probability \( P(T^- | D) \) stands for the probability that a false negative occurs while \( P(T^+ | D^c) \) stands for the probability that a false positive occurs. Obviously,

\[
P(T^+ | D) + P(T^- | D) = 1, \quad P(T^+ | D^c) + P(T^- | D^c) = 1.
\]

It follows from Bayes’ formula that

\[
P(D | T^+) = \frac{P(D)P(T^+ | D)}{P(D)P(T^+ | D) + P(D^c)P(T^+ | D^c)}
\]

(4.15)

and

\[
P(D^c | T^-) = \frac{P(D^c)P(T^- | D^c)}{P(D^c)P(T^- | D^c) + P(D)P(T^- | D)}.
\]

The former is called the predictive value of a positive test and the latter the one of a negative test.

For illustration, we set \( P(T^+ | D) = 0.7, \ P(T^- | D^c) = 0.99 \) and \( P(D) = d \). Note that \( P(D) \) is difficult to know in practical application. Substituting these probability into (4.15), we obtain the predictive value of a positive test:

\[
P(D | T^+) = \frac{0.7d}{0.7d + 0.01(1-d)} = \frac{70d}{1 + 69d}, \quad 0 \leq d \leq 1.
\]

(4.16)

Note that \( P(D | T^+) \) varies from 0 to 1. If \( P(D | T^+) \) is close to 1 the test is effective from the point of view of medical treatment. Thus, the formula (4.16) is useful for judging the performance of the test particularly when \( d \) is small, see Fig. 4.4.

Exercise 4.27. There are 10 lottery tickets with serial numbers from 1 to 10. Two tickets with numbers 1 and 2 are winning ones. A boy got 4 tickets.

(1) The boy says that he has a ticket of number 1. Find the probability that there is a winning ticket in the 6 rest tickets.

\[
\]

Advanc
(2) The boy says that he has at least one winning ticket. Find the probability that there is a winning ticket in the 6 rest tickets.

The above result would be counterintuitive.

Exercise 4.28. There are 100 patients in a hospital with a certain disease. Of these, 10 are selected to undergo a drug treatment that increases the percentage cured rate from 50% to 80%. Find the probability that the patient received a drug treatment if the patient is known to be cured.

5. Random Variables

5.1 Random variables and their distributions

A random variable is intuitively a variable whose values appear along with a certain probability law. A typical example appears in random sampling. Consider a variable whose values are obtained from the measurement of samples chosen randomly from a population. Then the variable obeys a certain probability law arising from random sampling, so it is a random variable.

To be slightly more precise, a random variable is a variable $X$ for which we may ask the probability $P(X \leq x)$ that $X$ takes values less than or equal $x \in \mathbb{R}$. However, for logical validity of $P(X \leq x)$ we need to prepare a probability space $(\Omega, \mathcal{F}, P)$ before introducing a random variable. In the above-mentioned example of random sampling, we set $\Omega$ to be the population and define the probability by $P(A) = |A|/|\Omega|$ along with combinatorial probability. Our variable $X$ gives a definite value for each individual $\omega$. In other words, $X : \Omega \rightarrow \mathbb{R}$ is a function. Then $P(X \leq x)$ is defined by

$$P(X \leq x) = \frac{|\{\omega \in \Omega; X(\omega) \leq x\}|}{|\Omega|}, \quad x \in \mathbb{R},$$  \hspace{1cm} (5.1)$$

where $\{X \leq x\}$ is a short-hand notation for $\{\omega \in \Omega; X(\omega) \leq x\}$. Abstractioning the above argument, we give the following formal definition.

**Definition 5.1.** Let $(\Omega, \mathcal{F}, P)$ be a probability space. A function $X : \Omega \rightarrow \mathbb{R}$ is called a random variable if $\{X \leq x\} = \{\omega \in \Omega; X(\omega) \leq x\}$ is an event in $\mathcal{F}$ for all $x \in \mathbb{R}$. Moreover, the function

$$F(x) = F_X(x) = P(X \leq x), \quad x \in \mathbb{R},$$  \hspace{1cm} (5.2)$$

is called the distribution function of $X$.

**Example 5.2.** Tossing a coin, we set $X = 1$ if the heads occurs and $X = 0$ if the tails occurs. Then $X$ becomes a random variable such that

$$P(X = 0) = P(X = 1) = \frac{1}{2}.$$  

The distribution function is given by

$$F_X(x) = \begin{cases} 
0, & x < 0, \\
1/2, & 0 \leq x < 1, \\
1, & x \geq 1.
\end{cases}$$

Rolling dice being similar, the investigation is left to the readers in the following exercise.
Exercise 5.3. Consider the experiment of rolling a fair die. Let $X$ be the random variable which assigns 1 if the number appears is even and 0 if the number that appears is odd. Find $P(X = 1)$ and $P(X = 0)$.

Exercise 5.4. Consider the experiment of tossing a coin three times. Let $X$ be the number of heads obtained. We assume that the tosses are independent and the probability of a head is $p$. Find the probabilities $P(X = 0), P(X = 1), P(X = 2)$ and $P(X = 3)$.

Exercise 5.5. Suppose that a fair die is rolled seven times. Find the probability that 1 and 2 dots appear twice each; 3, 4, and 5 dots once each; and 6 dots not at all.

Example 5.6. Let $\Omega$ be the set of players of Team A. Let $X$ be the height of a player randomly chosen from $\Omega$. Then the distribution function of $X$ is given by

$$F_X(x) = P(X \leq x) = \frac{|\{\omega \in \Omega; X(\omega) \leq x\}|}{|\Omega|}.$$ (5.3)

This is essentially the cumulative relative frequencies of the heights of Team A, see Sect. 2.1.

Theorem 5.7. Let $X$ be a random variable and $F(x) = F_X(x)$ the distribution function. Then we have

(i) $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to +\infty} F(x) = 1$;

(ii) If $x_1 \leq x_2$, then $F(x_1) \leq F(x_2)$;

(iii) $\lim_{x \to a^+} F(x) = F(a)$, namely, $F(x)$ is right-continuous.

Theorem 5.8. Let $X$ be a random variable and $F(x) = F_X(x)$ the distribution function. Then we have

$$P(X = x) = F_X(x) - \lim_{\epsilon \to 0^+} F_X(x - \epsilon), \quad x \in \mathbb{R}.$$ (5.4)

For the proofs see the standard textbooks. We only mention here that countable operations of sets are required in the proofs.

Exercise 5.9. Verify the properties (i)–(iii) in Theorem 5.7 for the distribution function in Example 5.2.

Definition 5.10. A random variable $X$ is called discrete if the distribution function $F_X(x)$ increases only by jumps. A random variable $X$ is called continuous if the distribution function $F_X(x)$ is continuous. (Note that there are random variables that are neither discrete nor continuous.)

For a discrete random variable $X$ the jump points of $F_X(x)$ are at most countable, say, $a_1, a_2, \ldots$. The jump at $x = a_i$ is denoted by $p_i > 0$. Then we have

$$p_i = P(X = a_i) = F_X(a_i) - \lim_{\epsilon \to 0^+} F_X(a_i - \epsilon), \quad \sum_i p_i = 1.$$ (5.5)

Thus, with a discrete random variable $X$, we may associate the possible values $a_i$ and its probability $p_i$. It is convenient to allow $p_i = 0$ (in that case $a_i$ is not a possible value though). The random variables in Examples 5.2 and 5.6 are discrete.

For a continuous random variable $X$, we have $P(X = x) = 0$ for all $x$. This is an immediate consequence of Theorem 5.8. We now understand why we needed to consider the probability of the events $\{X \leq x\}$ instead of $\{X = x\}$ for introducing a random variable.

Example 5.11. Let $X$ be the coordinate of a point chosen from an interval $\Omega = [0, L], L > 0$, in such a way that every point of $\Omega$ is chosen equally likely, see Example 4.4. We are interested in the distribution function $F_X(x)$. Since the event $\{X \leq x\}$ never occurs if $x < 0$, we have $F_X(x) = 0$ for $x < 0$. While the event $\{X \leq x\}$ certainly occurs if $x > L$, we have $F_X(x) = 1$ for $x > L$. For $0 \leq x \leq L$ we have

$$P(X \leq x) = \frac{|[0, x]|}{|[0, L]|} = \frac{x}{L}.$$ (5.6)

Consequently, we have

$$F_X(x) = \begin{cases} 0, & x < 0, \\ x/L, & 0 \leq x \leq L, \\ 1, & x > L. \end{cases}$$ (5.7)

Since $F_X(x)$ is a continuous function obviously, the random variable $X$ is continuous.

Example 5.12. Cutting off a stick of length $L$ at a randomly chosen point, we obtain two fragments. We are interested in the length of the shorter fragment, which is denoted by $S$. The stick is modeled by an interval $\Omega = [0, L]$ in the real line and let $X$ denote the coordinate of a randomly chosen point. Then $X$ becomes a random variable as is discussed in...
Example 5.11. Since $[S \leq x]$ never occurs for $x < 0$ and $[S \leq x]$ certainly occurs if $x > L/2$, we have $F_S(x) = 0$ for $x < 0$ and $F_S(x) = 1$ for $x > L/2$. Suppose that $0 \leq x \leq L/2$. Then we have

$$P(S \leq x) = P(0 \leq X \leq x) + P(L - x \leq X \leq L) = \frac{x}{L} + \frac{x}{L} = \frac{2x}{L}. $$

Summing up, we have

$$F_S(x) = \begin{cases} 
0, & x < 0, \\
2x/L, & 0 \leq x \leq L/2, \\
1, & x > L/2. 
\end{cases} \quad (5.5)$$

Thus, $S$ is a continuous random variable. We may derive the distribution function $F_S(x)$ alternatively by using $S = \min[X, L - X]$.

Example 5.13. Let $\Omega$ be a disc of radius $R > 0$. Choose a point randomly from $\Omega$ and let $X$ be the distance between the chosen point and the center of the disc. Then $X$ becomes a random variable. Obviously, $F_X(x) = 0$ for $x < 0$ and $F_X(x) = 1$ for $x > R$. Suppose that $0 \leq x \leq R$. Since the event $[X \leq x]$ corresponds to the concentric disc with radius $x$, we have

$$P(X \leq x) = \frac{\pi x^2}{\pi R^2} = \frac{x^2}{R^2},$$

where the probability is calculated along with Example 4.5. Consequently, we have

$$F_X(x) = \begin{cases} 
0, & x < 0, \\
x^2/R^2, & 0 \leq x \leq R, \\
1, & x > R. 
\end{cases} \quad (5.6)$$

Thus, $X$ is a continuous random variable.

In general, the distribution function $F_X(x)$ of a continuous random variable $X$ is continuous by definition but not necessarily differentiable. If $F_X(x)$ is piecewise differentiable, the derivative

$$f_X(x) = \frac{d}{dx} F_X(x)$$

is called the (probability) density function of $X$. It then follows from the fundamental theorem of differential integral calculus that

$$P(X \leq x) = F_X(x) = \int_{-\infty}^{x} f_X(t) dt$$

and

$$P(a \leq X \leq b) = \int_{a}^{b} f_X(x) dx, \quad a < b.$$ 

Since the density function gives a probability only through integration, if $F_X(x)$ is not differentiable at $x = a$, the value of $f_X(x)$ at $x = a$ may be given arbitrarily. Continuous random variables with density functions as well as discrete random variables cover a quite wide range of applications.

Exercise 5.14. Define a function $F(x)$ by

$$F(x) = \begin{cases} 
0, & x < 0, \\
x + \frac{1}{2}, & 0 \leq x \leq \frac{1}{2}, \\
1, & x \geq \frac{1}{2}. 
\end{cases}$$

Verify the properties (i)--(iii) in Theorem 5.7.

Exercise 5.15. Let $X$ be a random variable of which the distribution function is given by $F(x)$ described in Exercise 5.14. Find the following probabilities:

$$P\left( X \leq \frac{1}{4} \right), \quad P\left( 0 < X \leq \frac{1}{4} \right), \quad P(X = 0).$$

(Note that $F(x)$ is not continuous at $x = 0$.)

Exercise 5.16. Determine the constants $a$ and $b$ such that
\[ F(x) = \begin{cases} 1 - ae^{-x/b}, & x \geq 0, \\ 0, & x < 0, \end{cases} \]

is the distribution function of a random variable.

**Definition 5.17.** The **mean or expectation** of a random variable \( X \) is defined by

\[ \mathbb{E}[X] = \mu_X = \int_{\Omega} X(\omega)P(d\omega), \]

where the right-hand side is the so-called the Lebesgue integral.

For practical problems we consider two cases. For a discrete random variable \( X \) with possible values \( a_1, a_2, \ldots \) the mean becomes

\[ \mathbb{E}[X] = \mu_X = \sum_i a_i P(X = a_i) = \sum_i x P(X = x). \]

In the most right expression, which is just by convention, the sum is taken over all real numbers \( x \) but in fact, since \( P(X = x) = 0 \) except at most countable \( x = a_i \), the expression is reduced to a usual sum. For a continuous random variable with density function \( f_X(x) \) the mean becomes

\[ \mathbb{E}[X] = \mu_X = \int_{-\infty}^{+\infty} x f_X(x) dx. \]

Many important statistics of a random variable \( X \) is defined in terms of the mean. For example, the **variance** of \( X \) is defined by

\[ \text{Var}[X] = \sigma^2 = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2. \]

Moreover, the **central moment of degree** \( k \) is defined by

\[ m_k[X] = \mathbb{E}[(X - \mathbb{E}[X])^k]. \]

**Example 5.18.** Let \( X \) be the random variable introduced in Example 5.11. It is a continuous random variable since the distribution function \( F_X(x) \) in (5.4) is continuous. The density function \( f_X(x) \) is obtained by differentiating \( F_X(x) \) as follows:

\[ f_X(x) = \begin{cases} 1/L, & 0 \leq x \leq L, \\ 0, & \text{otherwise}. \end{cases} \]

Then the mean of \( X \) is given by

\[ \mathbb{E}[X] = \int_{-\infty}^{+\infty} x f_X(x) dx = \int_0^L x \frac{1}{L} dx = \frac{L}{2}. \]

Similarly, we have

\[ \mathbb{E}[X^2] = \int_{-\infty}^{+\infty} x^2 f_X(x) dx = \int_0^L x^2 \frac{1}{L} dx = \frac{L^2}{3}. \]

Hence the variance is given by

\[ \text{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{L^2}{3} - \left( \frac{L}{2} \right)^2 = \frac{L^2}{12}. \]

The probability distribution defined by the density function (5.7) is called the **uniform distribution** on \([0, L]\). Accordingly, the random variable \( S \) introduced in Example 5.12 obeys the uniform distribution on \([0, L/2]\).

**Example 5.19.** For \( \mu \in \mathbb{R} \) and \( \sigma > 0 \), the **normal distribution** \( N(\mu, \sigma^2) \) is defined by the density function:

\[ f(x) = \frac{1}{\sqrt{2\pi}\sigma^2} \exp\left[-\frac{(x - \mu)^2}{2\sigma^2}\right], \]

see also Example 2.19. In particular, \( N(0, 1) \) is called the **standard normal distribution**. If a random variable obeys \( N(0, 1) \), the distribution function is given by

\[ F_X(x) = P(X \leq x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt. \]

It is noted that the right-hand side is not expressed in terms of an elementary function. Instead, we define the **Gauss error function** by...
erf(x) = \frac{2}{\sqrt{\pi}} \int_{0}^{x} e^{-t^2} dt.

Then (5.8) becomes

\[ F_X(x) = \frac{1}{2} \left[ 1 + \text{erf} \left( \frac{x}{\sqrt{2}} \right) \right]. \]

**Exercise 5.20.** Let X be a discrete random variable such that

\[ P(X = -1) = P(X = 0) = P(X = 1) = \frac{1}{3}. \]

Find the mean and variance of X.

**Exercise 5.21.** Let X be a continuous random variable of which the density function is given by

\[ f_X(x) = \begin{cases} 2x, & 0 < x < 1, \\ 0, & \text{otherwise}. \end{cases} \]

Find the mean and variance of X.

**Exercise 5.22.** Let X be the random variable introduced in Example 5.13. Find the density function of X and show that E[X] = 2R/3 and V[X] = R^2/18.

**Exercise 5.23.** Prove that the moment of degree 2m of the standard normal distribution N(0, 1) is given by

\[ \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} x^{2m} e^{-x^2/2} dx = \frac{(2m)!}{2^{m}m!}, \quad m = 1, 2, \ldots. \]

### 5.2 Joint distributions

Let \( X_1, X_2, \ldots, X_n \) be random variables defined on a probability space \((\Omega, \mathcal{F}, P)\). There are two points of view. One is to regard them as a sequence of random variables. This is suitable for the study of asymptotic properties and limit behavior. The other is to regard them as a random vector \((X_1, X_2, \ldots, X_n)\) in \(n\)-dimensional space. Since the essence is the same, we switch the notation by convenience. The statistics of finitely many random variables \( X_1, X_2, \ldots, X_n \) is described by the joint distribution function defined by

\[ F_{X_1, X_2, \ldots, X_n}(x_1, x_2, \ldots, x_n) = P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n), \quad x_1, x_2, \ldots, x_n \in \mathbb{R}, \]

where the right-hand side is the probability of the product event \( \bigcap_{i=1}^{n} \{ X_i \leq x_i \} \).

If \( X_1, \ldots, X_n \) are discrete random variables, it is sufficient and more convenient to deal with the joint probability of the form

\[ P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n), \]

where \( x_1, x_2, \ldots, x_n \) run over all possible values of \( X_1, X_2, \ldots, X_n \), respectively. In that case the random points \( (X_1, X_2, \ldots, X_n) \) are scattered in \(n\)-dimensional space in a discrete manner. We are also interested in a particular type of continuous random vector, where the joint distribution function is given by the integral:

\[ F_{X_1, X_2, \ldots, X_n}(x_1, x_2, \ldots, x_n) = \int_{-\infty}^{x_1} dt_1 \int_{-\infty}^{x_2} dt_2 \cdots \int_{-\infty}^{x_n} dt_n f(t_1, t_2, \ldots, t_n), \]

for \( x_1, x_2, \ldots, x_n \in \mathbb{R} \). In that case, the integrand \( f(t_1, t_2, \ldots, t_n) \) is called the joint density function of \( X_1, X_2, \ldots, X_n \) and denoted by \( f_{X_1, X_2, \ldots, X_n}(x_1, x_2, \ldots, x_n) \).

**Exercise 5.24.** Consider an experiment of tossing a fair coin twice. Let \((X, Y)\) be a 2-dimensional random vector, where X is the number of heads that occurs in the two tosses and Y is the number of tails that occurs in the two tosses. Find \( P(X = 2, Y = 0) \), \( P(X = 0, Y = 1) \) and \( P(X = 1, Y = 1) \).

Let \((X_1, \ldots, X_n)\) be an \(n\)-dimensional random vector such that each \( X_j \) is a discrete random variable. Then we have

\[ P(X_1 = x) = \sum_{x_2, \ldots, x_n} P(X_1 = x, X_2 = x_2, \ldots, X_n = x_n), \]

and the mean of \( X_1 \) is given by

\[ \mu_{X_1} = E[X_1] = \sum_{x} x P(X_1 = x) = \sum_{x_1, x_2, \ldots, x_n} x_1 P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n). \]

Similarly,
The correlation coefficient

If \((X_1, \ldots, X_n)\) admits a joint density function \(f_{X_1, \ldots, X_n}(x_1, \ldots, x_n)\), we have

\[
f_{X_i}(x) = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} f_{X_1, \ldots, X_n}(x_1, \ldots, x_n)dx_2 \cdots dx_n,
\]
and the mean of \(X_1\) is given by

\[
\mu_{X_1} = \mathbb{E}[X_1] = \int_{-\infty}^{+\infty} xf_{X_1}(x)dx = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} x_1f_{X_1, \ldots, X_n}(x_1, \ldots, x_n)dx_1dx_2 \cdots dx_n.
\]

Similarly,

\[
\mu_{X_i} = \mathbb{E}[X_i] = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} x_jf_{X_i}(x)dx = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} x_1f_{X_i, \ldots, X_n}(x_1, \ldots, x_n)dx_1dx_2 \cdots dx_n.
\]

Moreover, the higher-order statistics are defined by means of \(\mathbb{E}[X_1^n \cdots X_n^n]\). For example, \(\mathbb{E}[X_1^2]\) is the moment of 2nd order and \(\mathbb{E}[X_1X_2]\) is a mixed moment of 2nd order. For discrete random variables we have

\[
\mathbb{E}[X_1X_2] = \sum_{x_1,x_2, \ldots, x_n} x_1x_2p(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)
\]
and for continuous random variables with a joint density function we have

\[
\mathbb{E}[X_1X_2] = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} x_1x_2f_{X_1X_2}(x_1, x_2, \ldots, x_n)dx_1dx_2 \cdots dx_n.
\]

**Definition 5.25.** The **covariance** of two random variables \(X\) and \(Y\) is defined by

\[
\sigma_{XY} = \text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].
\]

The **correlation coefficient** of \(X\) and \(Y\) is defined by

\[
\rho_{XY} = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}[X]\text{Var}[Y]}} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y},
\]
where \(\sigma_X = \sqrt{\text{Var}[X]}\) and \(\sigma_Y = \sqrt{\text{Var}[Y]}\) are the standard deviations of \(X\) and \(Y\), respectively.

For random variables \(X_1, \ldots, X_n\), the matrix \(\Sigma\) with

\[
\Sigma = [\sigma_{ij}], \quad \sigma_{ii} = \sigma_{X_i}^2 = \text{Var}[X_i], \quad \sigma_{ij} = \sigma_{X_iX_j} = \text{Cov}(X_i, X_j)
\]
is called the **variance-covariance matrix**.

**Definition 5.26.** We say that random variables \(X_1, X_2, \ldots, X_n\) are **independent** if the joint distribution function is factorized as

\[
P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n) = \prod_{i=1}^{n} P(X_i \leq x_i)
\]
or equivalently,

\[
F_{X_1, \ldots, X_n}(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} F_{X_i}(x_i).
\]

It is proved by definition that discrete random variables \(X_1, \ldots, X_n\) are independent if and only if

\[
P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \prod_{i=1}^{n} P(X_i = x_i)
\]
for all \(x_1, x_2, \ldots, x_n \in \mathbb{R}\). Random variables \(X_1, \ldots, X_n\) with joint density function \(f_{X_1, \ldots, X_n}(x_1, x_2, \ldots, x_n)\) are independent if and only if the joint density function is factorized as

\[
f_{X_1, \ldots, X_n}(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i),
\]
where \(f_{X_i}(x_i)\) is the density function of \(X_i\).

**Remark 5.27.** Definition 5.26 applies to an arbitrary family of random variables. A family of random variables \(\{X_1; \lambda \in \Lambda\}\) is called **independent** if any finitely many random variables \(X_{i_1}, \ldots, X_{i_n}\) chosen from the family are independent in the sense of Definition 5.26.
Remark 5.28. By definition two random variables $X$ and $Y$ are independent if $P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y)$ for all $x, y \in \mathbb{R}$. A family of random variables $\{X_\lambda: \lambda \in \Lambda\}$ is called pairwise independent if any two random variables $X_{\lambda_1}$ and $X_{\lambda_2}$, $\lambda_1 \neq \lambda_2$, chosen from the family are independent. Note that a pairwise independent family of random variables is not necessarily independent.

Exercise 5.29. For $\alpha > 0$ and $\beta > 0$ let $F(x, y)$ be a function defined by

$$F(x, y) = \begin{cases} (1 - e^{-\alpha x})(1 - e^{-\beta y}), & x \geq 0, y \geq 0, \\ 0, & \text{otherwise}. \end{cases}$$

Prove that $F(x, y)$ is the joint distribution function of a 2-dimensional random vector $(X, Y)$. Then show that $X$ and $Y$ are independent.

Theorem 5.30. If two random variables $X$ and $Y$ are independent, we have $E[XY] = E[X]E[Y]$ and $\text{Cov}(X, Y) = 0$.

Proof. Suppose that $X$ and $Y$ are discrete random variables. Since they are independent by assumption, we have the factorization $P(X = x, Y = y) = P(X = x)P(Y = y)$. Then we have

$$E[XY] = \sum_{x,y} xyP(X = x, Y = y) = \sum_{x} xP(X = x) \sum_{y} yP(Y = y)$$

and hence

$$E[XY] = E[X]E[Y]. \quad (5.11)$$

Suppose next that $X$ and $Y$ admits a joint density function $f_{XY}(x, y)$. Since they are independent by assumption we have $f_{XY}(x, y) = f_X(x)f_Y(y)$. Then we have

$$E[XY] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xyf_{XY}(x, y)dx dy = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xyf_X(x)f_Y(y)dx dy = \int_{-\infty}^{\infty} f_X(x)dx \int_{-\infty}^{\infty} f_Y(y)dy$$

and we come to (5.11). For a general pair of independent random variables $X$ and $Y$ we need Lebesgue integral on a probability space and omit the proof, see the standard textbooks. Finally, it follows immediately from (5.11) that

$$\text{Cov}(X, Y) = E[XY] - E[X]E[Y] = E[X]E[Y] - E[X]E[Y] = 0,$$

as desired. \qed

Remark 5.31. Two random variables $X$ and $Y$ are called uncorrelated if $\text{Cov}(X, Y) = 0$. Theorem 5.30 says that independent random variables are uncorrelated. However, the converse is not true in general, see Exercise 5.32.

Exercise 5.32. Let $Z_1$ and $Z_2$ be independent random variables such that

$$P(Z_1 = \pm 1) = P(Z_2 = \pm 1) = \frac{1}{2},$$

in other words, $Z_1$ and $Z_2$ stand for tossing two coins. Set

$$X = Z_1 + Z_2, \quad Y = Z_1 - Z_2.$$

Show that $X$ and $Y$ are uncorrelated but are not independent.

Exercise 5.33. Let $(X, Y)$ be a 2-dimensional random vector of which the density function is given by

$$f_{XY}(x, y) = \frac{x^2 + y^2}{4\pi} e^{-(x^2+y^2)/2}.$$

Show that $X$ and $Y$ are uncorrelated but are not independent.

5.3 Regression curves

Let $X$ and $Y$ be two random variables. We identify $(X, Y)$ with a random point in the $xy$-coordinate plane. First we consider the case where both $X$ and $Y$ are discrete. For $x, y \in \mathbb{R}$ the conditional probability

$$P(Y = y|X = x) = \frac{P(X = x, Y = y)}{P(X = x)}$$

is defined whenever $P(X = x) > 0$. Note that

$$\sum_y P(Y = y|X = x) = \frac{1}{P(X = x)} \sum_y P(X = x, Y = y) = \frac{1}{P(X = x)} \times P(X = x) = 1.$$

Then we regard $P(Y = y|X = x)$ as a probability distribution concentrated on the vertical line with $x$-coordinate $x$ in the $xy$-coordinate plane. Then the conditional expectation of $Y$ under the condition $X = x$ is defined by
Then we obtain a function \( x \mapsto \mathbb{E}[Y|X = x] \), where \( x \) runs over \( \mathbb{R} \) such that \( P(X = x) > 0 \). This function gives rise to a discrete curve in the \( xy \)-coordinate plane, which is called the regression curve for \( Y \) subject to \( X \).

We next consider the case where \( X \) and \( Y \) admit a joint density function \( f_{XY}(x, y) \). The conditional density function of \( Y \) under the condition \( X = x \) is defined by

\[
   f_{Y|X}(y|x) = \frac{f_{XY}(x, y)}{f_X(x)} = \frac{1}{\int_{-\infty}^{\infty} f_{XY}(x, y) dy},
\]

whenever the denominator is positive. Since we understand that \( f_{Y|X}(y|x) \) is a density function concentrated on the vertical line with \( x \)-coordinate is \( x \) in the \( xy \)-coordinate plane. Then the conditional expectation of \( Y \) under the condition \( X = x \) is defined by

\[
   \mathbb{E}[Y|X = x] = \int_{-\infty}^{\infty} y f_{Y|X}(y|x) dy.
\]

5.4 Two-dimensional normal distributions

Let \( \mu = [\mu_j] \) be an \( n \)-dimensional column vector and \( \Sigma = [\sigma_{jk}] \) a strictly positive definite \( n \times n \) matrix. By definition \( (x, \Sigma x) > 0 \) for all \( x \in \mathbb{R}^n \) with \( x \neq 0 \) and necessarily \( \Sigma \) is invertible and symmetric. Define a function \( f(x) \) by

\[
   f(x) = \frac{1}{\sqrt{(2\pi)^n|\Sigma|}} \exp \left[ -\frac{1}{2} (x - \mu)' \Sigma^{-1} (x - \mu) \right], \quad x \in \mathbb{R}^n,
\]

where \( |\Sigma| \) is the determinant. It is proved that

\[
   \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \ldots, x_n) dx_1 \cdots dx_n = 1,
\]

with the help of diagonalization of \( \Sigma \) and coordinate change. In other words, \( f(x) \) is a probability density function in \( n \) variables. The corresponding probability distribution is called an \( n \)-dimensional normal distribution and is denoted by \( \mathcal{N}(\mu, \Sigma) \). Moreover, we can check by elementary calculus that

\[
   \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} x_j f(x_1, \ldots, x_n) dx_1 \cdots dx_n = \mu_j
\]

and

\[
   \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} (x_j - \mu_j)(x_k - \mu_k) f(x_1, \ldots, x_n) dx_1 \cdots dx_n = \sigma_{jk}.
\]

As a result, \( \mu \) is the mean vector and \( \Sigma \) the variance-covariance matrix of the normal distribution \( \mathcal{N}(\mu, \Sigma) \).

Here we study the case of two dimension. Take a vector \( \mu \in \mathbb{R}^2 \) and a strictly positive definite \( 2 \times 2 \) matrix \( \Sigma \), say,
\[ \mathbf{\mu} = \begin{bmatrix} a \\ b \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}. \]

Note that \( \Sigma \) becomes a symmetric matrix i.e., \( \sigma_{12} = \sigma_{21} \). The density function of \( N(\mathbf{\mu}, \Sigma) \) is defined by

\[ f(x, y) = \frac{1}{\sqrt{2\pi|\Sigma|}} \exp\left( -\frac{1}{2} ((x - \mu, \Sigma^{-1}(x - \mu)) \right), \quad x = \begin{bmatrix} x \\ y \end{bmatrix} \in \mathbb{R}^2. \quad (5.18) \]

\[ h(x, y) = \frac{1}{2\pi \sigma_X \sigma_Y \sqrt{1-\rho^2}} \exp\left( -\frac{1}{2(1-\rho^2)} \left\{ \left( \frac{x - \mu_X}{\sigma_X} \right)^2 - 2\rho \frac{x - \mu_X}{\sigma_X} \frac{y - \mu_Y}{\sigma_Y} + \left( \frac{y - \mu_Y}{\sigma_Y} \right)^2 \right\} \right). \quad (5.19) \]

**Proof.** Let \( N(\mathbf{\mu}, \Sigma) \) be the normal distribution that \( (X, Y) \) obeys and \( f(x, y) \) its density function as in (5.18). As a particular case of (5.16) and (5.17), we have

\[ \mu_X = E[X] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} xf(x, y)dx dy = a, \]

\[ \mu_Y = E[Y] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} yf(x, y)dx dy = b, \]

and

\[ \sigma_X^2 = E[(X - \mu_X)^2] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (x - \mu_X)^2 f(x, y)dx dy = \sigma_{11}, \]

\[ \sigma_Y^2 = E[(Y - \mu_Y)^2] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (y - \mu_Y)^2 f(x, y)dx dy = \sigma_{22}, \]

\[ \sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (x - \mu_X)(y - \mu_Y)f(x, y)dx dy = \sigma_{12} = \sigma_{21}. \]

Hence the joint density function \( f_{XY}(x, y) \) is given by the density function of \( N(\mathbf{\mu}, \Sigma) \) with \( \mu \) and \( \sigma \) being given as above. We now look at the quadratic function \( \langle(x - \mu), \Sigma^{-1}(x - \mu) \rangle \) in the right-hand side of (5.18). First setting

\[ \Sigma^{-1} = \begin{bmatrix} \tau_{11} & \tau_{12} \\ \tau_{21} & \tau_{22} \end{bmatrix}, \]

we obtain

\[ \langle(x - \mu), \Sigma^{-1}(x - \mu) \rangle = \tau_{11}(x - a)^2 + 2\tau_{12}(x - a)(y - b) + \tau_{22}(y - b)^2. \quad (5.20) \]
Then inserting

\[ \tau_{11} = \frac{\sigma_{22}}{|\Sigma|}, \quad \tau_{22} = \frac{\sigma_{11}}{|\Sigma|}, \quad \tau_{12} = \tau_{21} = -\frac{\sigma_{12}}{|\Sigma|}, \]

into (5.20), we have

\[ \langle (x - \mu), \Sigma^{-1}(x - \mu) \rangle = \frac{1}{|\Sigma|} \left\{ \sigma_{22}(x - a)^2 - 2\sigma_{12}(x - a)(y - b) + \sigma_{11}(y - b)^2 \right\} = \sigma_{11}\sigma_{22} \left\{ \frac{(x - a)^2}{\sigma_{11}} - 2\frac{\sigma_{12}}{\sigma_{11}\sigma_{22}}(x - a)(y - b) + \frac{(y - b)^2}{\sigma_{22}} \right\}. \]  

Finally, using the correlation coefficient

\[ \rho = \frac{\sigma_{XY}}{\sigma_X\sigma_Y} = \frac{\sigma_{12}}{\sqrt{\sigma_{11}\sigma_{22}}} \]

together with \( a = \mu_X, b = \mu_Y \), we come to

\[ \langle (x - \mu), \Sigma^{-1}(x - \mu) \rangle = \frac{\sigma_{22}^2}{|\Sigma|} \left\{ \frac{(x - \mu_x)^2}{\sigma_X^2} - 2\rho \frac{x - \mu_x y - \mu_y}{\sigma_X \sigma_Y} + \left( \frac{y - \mu_y}{\sigma_Y} \right)^2 \right\}. \]  

On the other hand, we have

\[ |\Sigma| = \sigma_{11}\sigma_{22} - \sigma_{12}\sigma_{21} = \sigma_X^2\sigma_Y^2 - \sigma_{XY}^2 \sigma_X^2 \sigma_Y^2 \left( 1 - \frac{\sigma_{XY}^2}{\sigma_X^2 \sigma_Y^2} \right) = \sigma_X^2 \sigma_Y^2 (1 - \rho^2). \]

Then (5.19) follows immediately from (5.22) and (5.23).

**Theorem 5.37.** Let \( X \) and \( Y \) be random variables with means \( \mu_X, \mu_Y \), variances \( \sigma_X^2, \sigma_Y^2 \) and covariance \( \sigma_{XY} \). If \( (X, Y) \) obeys a 2-dimensional normal distribution, the density functions of \( X \) and \( Y \) (called the marginal density function in this context) are given by

\[ f_X(x) = \int_{-\infty}^{+\infty} f_{XY}(x, y) dy = \frac{1}{\sqrt{2\pi\sigma_X^2}} \exp \left[ -\frac{(x - \mu_X)^2}{2\sigma_X^2} \right], \]  

\[ f_Y(y) = \int_{-\infty}^{+\infty} f_{XY}(x, y) dx = \frac{1}{\sqrt{2\pi\sigma_Y^2}} \exp \left[ -\frac{(y - \mu_Y)^2}{2\sigma_Y^2} \right], \]

respectively. In other words, \( X \) and \( Y \) obey the normal distributions \( N(\mu_X, \sigma_X^2) \) and \( N(\mu_Y, \sigma_Y^2) \), respectively.

**Proof.** We see from (5.21) that

\[ \langle (x - \mu), \Sigma^{-1}(x - \mu) \rangle = \sigma_{11} \left( y - b - \frac{\sigma_{12}}{\sigma_{11}}(x - a) \right)^2 + \frac{1}{\sigma_{11}} (x - a)^2, \]

and hence

\[ f_{XY}(x, y) = \frac{1}{\sqrt{2\pi\sigma_X^2|\Sigma|}} \exp \left[ -\frac{1}{2} \frac{\sigma_{11}}{|\Sigma|} \left( y - b - \frac{\sigma_{12}}{\sigma_{11}}(x - a) \right)^2 \right]. \]

Then we have

\[ f_X(x) = \int_{-\infty}^{+\infty} f_{XY}(x, y) dy = \frac{1}{\sqrt{2\pi\sigma_X^2|\Sigma|}} \sqrt{2\pi|\Sigma|} \exp \left[ -\frac{1}{2\sigma_{11}} (x - a)^2 \right] \]

\[ = \frac{1}{\sqrt{2\pi\sigma_X^2}} \exp \left[ -\frac{1}{2\sigma_{11}} (x - a)^2 \right], \]

which proves (5.24). Similarly, (5.25) is derived.

**Theorem 5.38.** Let \( X \) and \( Y \) be random variables with means \( \mu_X, \mu_Y \), variances \( \sigma_X^2, \sigma_Y^2 \) and covariance \( \sigma_{XY} \). If \( (X, Y) \) obeys a 2-dimensional normal distribution, the conditional density function \( f_{Y|X}(y|x) \) is given by

\[ f_{Y|X}(y|x) = \frac{f_{XY}(x, y)}{f_X(x)} = \frac{1}{\sqrt{2\pi\sigma_Y^2(1 - \rho^2)}} \exp \left[ -\frac{1}{2\sigma_Y^2(1 - \rho^2)} \left( y - \mu_Y - \frac{\sigma_{XY}}{\sigma_Y^2} (x - \mu_X) \right)^2 \right], \]

where \( \rho = \rho_{XY} \) is the correlation coefficient of \( X \) and \( Y \). In particular, the conditional density function \( f_{Y|X}(y|x) \) is a normal distribution.

**Proof.** By taking the ratio of (5.26) against (5.27) we obtain
\[
f_{Y|X}(y|x) = \frac{f_{XY}(x, y)}{f_X(x)} = \frac{1}{\sqrt{2\pi\sigma_Y^2} |\Sigma|} \exp \left[ -\frac{1}{2\sigma_Y^2 |\Sigma|} \left( y - \frac{\sigma_{XY}}{\sigma_X} (x - \mu_X) \right)^2 \right].
\]

Then, using \(\sigma_{11}^{-1} |\Sigma| = \sigma_Y^2 (1 - \rho^2)\) we obtain (5.28).

From (5.28) we obtain immediately
\[
E[Y|X = x] = \int_{-\infty}^{+\infty} y f_{Y|X}(y|x) dy = \mu_Y + \frac{\sigma_{XY}}{\sigma_X} (x - \mu_X).
\]
Thus the regression curve for 2-dimensional normal distribution is a line determined by
\[
y = \mu_Y + \frac{\sigma_{XY}}{\sigma_X} (x - \mu_X),
\]
or equivalently by
\[
\frac{y - \mu_Y}{\sigma_Y} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \frac{x - \mu_X}{\sigma_X} = \rho_{XY} \frac{x - \mu_X}{\sigma_X},
\]  
(5.29)

where \(\rho_{XY}\) is the correlation coefficient of \(X\) and \(Y\).

In Sect. 3.4 we discussed the method of least square for a regression line. We apply a similar idea to joint density function. Namely, given a joint density function \(f_{XY}(x, y)\), we ask for a line \(y = ax + \beta\) which gives the best approximation. For each fixed \(x\) the fluctuation of \(Y\) under the condition \(X = x\) is measured by the so-called square error \(E[(Y - (ax + \beta))^2|X = x]\). The method of least square is to find the coefficients \(a\) and \(\beta\) in such a way that the total of square errors
\[
Q = \int_{-\infty}^{+\infty} E[(Y - (ax + \beta))^2|X = x] f_X(x) dx
\]  
(5.30)
is minimized. In fact, we have
\[
Q = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (y - (ax + \beta))^2 f_{XY}(y|x) f_X(x) dy dx
\]
\[
= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (y - (ax + \beta))^2 f_{XY}(x, y) dy dx
\]
\[
= E[(Y - (aX + \beta))^2].
\]

Note that \(Q = Q(a, \beta)\) is a quadratic function. Then, after simple calculus we see that the minimum of \(Q\) is attained at
\[
a = \frac{\sigma_{XY}}{\sigma_X^2}, \quad \beta = \mu_Y - \frac{\sigma_{XY}}{\sigma_X} \mu_X.
\]
Hence the regression line is given by
\[
y = \frac{\sigma_{XY}}{\sigma_X^2} x + \mu_Y - \frac{\sigma_{XY}}{\sigma_X} \mu_X,
\]
or equivalently by
\[
\frac{y - \mu_Y}{\sigma_Y} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \frac{x - \mu_X}{\sigma_X} = \rho_{XY} \frac{x - \mu_X}{\sigma_X},
\]  
(5.31)

Note that (5.29) and (5.31) coincide. Thus, we come to the following assertion.

**Theorem 5.39.** Let \(X\) and \(Y\) be two random variables and assume that \((X, Y)\) obeys a 2-dimensional normal distribution. Then the regression curve defined by \(x \mapsto E[Y|X = x]\) is given by
\[
y = \frac{\sigma_{XY}}{\sigma_X^2} x + \mu_Y - \frac{\sigma_{XY}}{\sigma_X} \mu_X,
\]
which coincides with the regression line determined by the method of least squares. In other words, the above line minimizes the total of square errors \(Q\) defined in (5.30).

**Remark 5.40.** For general distribution, the regression curve defined by \(x \mapsto E[Y|X = x]\) is not necessarily a line. Hence the regression line obtained by the method of least squares does not necessarily coincide with the regression curve determined by the conditional expectation.

**Exercise 5.41.** Let \((X, Y)\) be a 2-dimensional random vector of which the density function is given by
\[ f_{XY}(x, y) = \frac{1}{2\sqrt{3\pi}} \exp \left[ -\frac{1}{3} \left( x^2 - xy + y^2 + x - 2y + 1 \right) \right]. \]

1. Find the means of \( X \) and \( Y \).
2. Find the variances of \( X \) and \( Y \).
3. Find the correlation coefficient of \( X \) and \( Y \).

Exercise 5.42. Assume that a random vector \((X, Y)\) obeys a 2-dimensional normal distribution \( N(\mu, \Sigma) \). Show that \( \text{Cov}(X, Y) = 0 \) implies that \( X \) and \( Y \) are independent. (See Remark 5.31.)

Bibliographical Notes

There are many excellent textbooks on multi-variate analysis. Here we only mention some of them. For general introduction to multi-variate analysis, we refer to Anderson [1], also see Dobson–Barnett [7] for a new approach by generalized linear models. Wooldridge [21] contains basics of multi-variate analysis and further topics in econometrics. Hair–Black–Babin–Anderson [9] provides a comprehensive guideline of multi-variate analysis from a practical point of view. For introduction to statistics see Brink [4], Rumsey [18, 19], see also Hoel [12]. For sampling theory see e.g., Hansen–Hurwitz–Madow [11], Kish [14]. For Bayesian analysis see Robert [17], Smith [20]. For general introduction to probability theory we refer to Chung [6], Durrett [8], while Kolmogorov [15] is the origin of modern probability theory. For measure theoretical probability theory, where the measure space and Lebesgue integrals are crucial, see Athreya–Lahiri [2], Billingsley [3] and Šapiški-Kopp [5]. The old book Halmos [10] is also widely known. For more exercise on probability and random variables see Hsu [13]. For more on basics of data science, see e.g., Kotu–Desphande [16].

Acknowledgments

These lecture notes are based on the lectures delivered as a part of Data Science Basics in Graduate Program in Data Science (GP-DS) and Data Sciences Program (DSP) at Tohoku University in 2018–2020. The authors are grateful to the teaching stuffs for their invaluable assistance and constant encouragement. They thank the anonymous referees for their comments and suggestions that improved the presentation of these lecture notes.

REFERENCES

[1] Anderson, T. W., An Introduction to Multivariate Statistical Analysis, 3rd Ed., Wiley-Interscience, Hoboken, NJ (2003).
[2] Athreya, K. B., and Lahiri, S. N., Measure Theory and Probability Theory, Springer, New York (2006).
[3] Billingsley, P., Probability and Measure, Wiley Series in Probability and Statistics, John Wiley & Sons, Inc., Hoboken, NJ (2012).
[4] Brink, D., Statistics, Ventus Publishing ApS (2010).
[5] Šapiški, M., and Kopp, E., Measure, Integral and Probability, 2nd Ed., Springer-Verlag, London (2004).
[6] Chung, K. L., A Course in Probability Theory, 3rd Ed., Academic Press, Inc., San Diego, CA (2001).
[7] Dobson, A. J., and Barnett, A. G., An Introduction to Generalized Linear Models, 4th Ed., CRC Press, Boca Raton, FL (2018).
[8] Durrett, R., Probability — Theory and Examples, 5th Ed., Cambridge University Press, Cambridge (2019).
[9] Hair, J., Black, W., Babin, B., and Anderson, R., Multivariate Data Analysis, Cengage Learning EMEA (2018).
[10] Halmos, P. R., Measure Theory, D. Van Nostrand Company, Inc., New York, N.Y. (1950).
[11] Hansen, M. H., Hurwitz, W. N., and Madow, W. G., Sample Survey Methods and Theory, Wiley Classics Library, John Wiley & Sons, Inc., New York (1993).
[12] Hoel, P. G., Introduction to Mathematical Statistics, 5th Ed., Wiley (1984).
[13] Hsu, H. P., Schaum’s Outline of Probability, Random Variables, and Random Processes, 4th Ed., McGraw-Hill Education (2020).
[14] Kish, L., Survey Sampling, Wiley-Interscience (1995).
[15] Kolmogorov, A., Grundbegriffe der Wahrscheinlichkeitsrechnung, Julius Springer, Berlin (1933).
[16] Kotu, V., and Desphande, B., Data Science: Concepts and Practice, 2nd Ed., Morgan Kaufmann (2019).
[17] Robert, C., The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation, 2nd Ed., Springer (2007).
[18] Rumsey, D. J., Statistics for Dummies, 2nd Ed., Wiley Publishing (2011).
[19] Rumsey, D. J., Statistics Essentials for Dummies, John Wiley & Sons, Inc. (2019).
[20] Smith, J. Q., Bayesian Decision Analysis: Principles and Practice, Cambridge University Press (2010).
[21] Wooldridge, J. M., Econometric Analysis of Cross Section and Panel Data, 2nd Ed., MIT Press (2010).

Appendix A: Some Tips for Matrices and Vectors

A.1 Definition

A rectangular array of numbers is called a matrix, where each horizontal array is called a row and each vertical one a column. A matrix having \( n \) rows and \( p \) columns is called an \( n \times p \) matrix and is of the form:
As a rule, the rows are numbered from top to bottom and the columns are numbered from left to right. An entry of a matrix $X$ at the crossroad of the $i$th row and $j$th column is called the $(i, j)$-entry and is denoted by $(X)_{ij}$. For $X$ in (A.1) we have $(X)_{ij} = x_{ij}$.

As a special case, a matrix with a single column, that is, an $n \times 1$ matrix:

$$
\begin{bmatrix}
    x_1 \\
    x_2 \\
    \vdots \\
    x_n
\end{bmatrix}
$$

is called an $n$-dimensional column vector. A matrix with a single row, that is, a $1 \times n$ matrix

$$
\begin{bmatrix}
    x_1 & x_2 & \cdots & x_n
\end{bmatrix}
$$

is called an $n$-dimensional row vector. These vectors represent a point in $n$-dimensional coordinate space, where the choice of a column or row vector is up to the context.

We say that two matrices are of the same type or of the same size if the numbers of their columns coincide as well as that of their rows. For two matrices $X$ and $Y$ we say that $X = Y$ if they are of the same type and $(X)_{ij} = (Y)_{ij}$ for all $i$ and $j$. Note that two matrices are never equal if they are not of the same type.

### A.2 Transposition

For a matrix $X$ the transposed matrix, denoted by $X^T$, is defined by exchanging the rows and columns of $X$. If $X$ is an $n \times p$ matrix, the transposed matrix $X^T$ becomes a $p \times n$ matrix. Writing down their entries explicitly, we have

$$
X = 
\begin{bmatrix}
    x_{11} & \cdots & x_{ij} & \cdots & x_{ip} \\
    \vdots & \ddots & \vdots & \ddots & \vdots \\
    x_{1i} & \cdots & x_{ij} & \cdots & x_{ip} \\
    \vdots & \ddots & \vdots & \ddots & \vdots \\
    x_{ni} & \cdots & x_{nj} & \cdots & x_{np}
\end{bmatrix},
$$

$X^T = 
\begin{bmatrix}
    x_{11} & x_{1j} & \cdots & x_{i1} & \cdots & x_{ip} \\
    \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
    x_{1j} & \cdots & x_{ij} & \cdots & x_{nj} \\
    \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
    x_{nj} & \cdots & x_{nj} & \cdots & x_{np}
\end{bmatrix},
$$

where $x_{ij}$ is the $(i, j)$-entry of $X$ and it is the $(j, i)$-entry of $X^T$. In short,

$$(X^T)_{ij} = (X)_{ji}.$$ 

By repeating the operation of transposition twice, a matrix turns back to the original one. Namely, we have $(X^T)^T = X$.

Applying transposition, a column vector becomes a row vector and vice versa as follows:

$$
\begin{bmatrix}
    x_1 \\
    x_2 \\
    \vdots \\
    x_n
\end{bmatrix}^T = [x_1 \ x_2 \ \cdots \ x_n], \quad [x_1 \ x_2 \ \cdots \ x_n]^T =
\begin{bmatrix}
    x_1 \\
    x_2 \\
    \vdots \\
    x_n
\end{bmatrix}.
$$

It is often convenient to write a column vector as $[x_1 \ x_2 \ \cdots \ x_n]^T$ for saving space.

### A.3 Addition and scalar multiplication

Let $X$ and $Y$ be matrices of the same type, say,
Moreover, letting
\[ C_0 \]
In fact, for the subtraction we have
\[ O \]
Let
\[ M \]
away, but the real power of matrix multiplication will be understood after some patient practice. Let
\[ X \]
matrices, and assume that the number of columns of
\[ X \]
a
\[ p \]
\[ X = \begin{bmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nj} & \cdots & x_{np} \end{bmatrix}, \quad Y = \begin{bmatrix} y_{11} & \cdots & y_{1j} & \cdots & y_{1p} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ y_{i1} & \cdots & y_{ij} & \cdots & y_{ip} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ y_{n1} & \cdots & y_{nj} & \cdots & y_{np} \end{bmatrix} \]

Then their sum \( X + Y \) is defined entrywise, namely,
\[ X + Y = \begin{bmatrix} x_{11} + y_{11} & \cdots & x_{ij} + y_{ij} & \cdots & x_{1p} + y_{1p} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i1} + y_{i1} & \cdots & x_{ij} + y_{ij} & \cdots & x_{ip} + y_{ip} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{n1} + y_{n1} & \cdots & x_{nj} + y_{nj} & \cdots & x_{np} + y_{np} \end{bmatrix}. \]

The difference \( X - Y \) is similarly defined. These rules are written in a simpler form:
\[ (X \pm Y)_{ij} = (X)_{ij} \pm (Y)_{ij} = x_{ij} \pm y_{ij}. \]

Next, for a real number \( a \) the scalar multiplication \( aX \) is defined by
\[ aX = \begin{bmatrix} ax_{11} & \cdots & ax_{ij} & \cdots & ax_{1p} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ ax_{i1} & \cdots & ax_{ij} & \cdots & ax_{ip} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ ax_{n1} & \cdots & ax_{nj} & \cdots & ax_{np} \end{bmatrix}, \]

or equivalently,
\[ (aX)_{ij} = a(X)_{ij} = ax_{ij}. \]

Since the addition and scalar multiplication are defined entrywise, the usual calculation rules are valid also for matrices:
(i) \( X + Y = Y + X \);
(ii) \( X + (Y + Z) = (X + Y) + Z \);
(iii) \( a(X + Y) = aX + aY \);
(iv) \( (a + b)X = aX + bX \);
(v) \( (ab)X = a(bX) \).

Let \( O \) be the matrix with entries being all zero. If \( X \) and \( O \) are of the same type, we have
\[ X + O = O + X = X. \]

Moreover, letting \( -X = (-1)X \) (scalar multiplication by \(-1\)) we have
\[ X + (-X) = (-X) + X = O. \]

In fact, for the subtraction we have
\[ X - Y = X + (-Y) = X + (-1)Y. \]

Next we review the multiplication of matrices. The seemingly complicated definition likely keeps the beginners away, but the real power of matrix multiplication will be understood after some patient practice. Let \( X \) and \( Y \) be two matrices, and assume that the number of columns of \( X \) and that of rows of \( Y \) coincide, say, \( X \) is an \( n \times p \) matrix and \( Y \) is a \( p \times m \) matrix. The product \( XY \) is defined to be an \( n \times m \) matrix whose \((i, j)\)-entry is given by
\[ (XY)_{ij} = x_{i1}y_{1j} + \cdots + x_{ip}y_{pj}, \]

that is,
\[ (XY)_{ij} = \sum_{k=1}^{p} x_{ik}y_{kj} = \sum_{k=1}^{p} (X)_{ik}(Y)_{kj}. \quad (A.4) \]

In fact, \((XY)_{ij}\) is obtained from the \(i\)th row of \( X \) and the \(j\)th column of \( Y \):
\[
X = \begin{bmatrix}
  x_{11} & \cdots & x_{1k} & \cdots & x_{1p} \\
  \vdots & \ddots & \vdots & \ddots & \vdots \\
  x_{n1} & \cdots & x_{nk} & \cdots & x_{np}
\end{bmatrix}, \quad \quad \quad
Y = \begin{bmatrix}
  y_{11} & \cdots & y_{1j} & \cdots & y_{1m} \\
  \vdots & \ddots & \vdots & \ddots & \vdots \\
  y_{k1} & \cdots & y_{kj} & \cdots & y_{km} \\
  \vdots & \ddots & \vdots & \ddots & \vdots \\
  y_{p1} & \cdots & y_{pj} & \cdots & y_{pm}
\end{bmatrix}.
\]

It is instructive to note in the formula (A.4) that the sum is taken over \(k\) appearing as if connecting two matrices \(X\) and \(Y\). The other indices \(i\) and \(j\) are outside the summation and stand for the row and column numbers of the new matrix \(XY\).

By calculation we can check the associativity:

\[(XY)Z = X(YZ)\]

just as in the case of numbers. Thanks to the associativity we may write \(XYZ\) for the multiplication of three matrices without brackets. On the other hand, the commutativity does not hold, that is, we have \(XY \neq YX\) in general. First of all, if \(X\) and \(Y\) do not satisfy the condition on sizes, it happens that \(XY\) is defined but \(YX\) not. Even if both \(XY\) and \(YX\) are defined, their sizes do not necessarily coincide. Even if both \(XY\) and \(YX\) are defined and their sizes coincide, the entries do not necessarily coincide. As a matter of fact, two matrices \(X\) and \(Y\) are in a very special relation when \(XY = YX\) holds.

The \(n \times n\) matrix with diagonal entries are all one and the off-diagonal ones zero is called the identity matrix and is denoted by \(I\):

\[
I = \begin{bmatrix}
  1 & \cdots & 0 & \cdots & 0 \\
  \vdots & \ddots & \vdots & \ddots & \vdots \\
  0 & \cdots & 1 & \cdots & 0 \\
  \vdots & \ddots & \vdots & \ddots & \vdots \\
  0 & \cdots & 0 & \cdots & 1
\end{bmatrix}
\]

Then for any \(n \times n\) matrix \(X\) we have \(XI = IX = X\), namely \(I\) is the multiplication unit. If two \(n \times n\) matrices \(X\) and \(Y\) satisfy \(XY = I\), we have \(YX = I\). In that case we say that \(Y\) is the inverse matrix of \(X\) and is denoted by \(X^{-1}\). It is noted that the inverse matrix \(X^{-1}\) does not necessarily exist, but if exists it is uniquely determined. A simple application of inverse matrix makes us possible to write down the unique solution to an equation \(AX = B\) as \(X = A^{-1}B\) whenever the inverse matrix of \(A\) exists.

### A.4 Inner product and norm

The inner product of \(n\)-dimensional column vectors

\[
x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad \quad \quad
y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}
\]

is defined by

\[
\langle x, y \rangle = \sum_{i=1}^{n} x_i y_i. \quad (A.6)
\]

For the inner product the symbol \(x \cdot y\) is also used in literatures. According to the definition of matrix multiplication \(x^T y\) becomes a \(1 \times 1\) matrix with a single entry given by the right-hand side of (A.6). Then, it is convenient to identify the \(1 \times 1\) matrix \(x^T y\) with the single number \(\langle x, y \rangle\) and write

\[
\langle x, y \rangle = x^T y = \sum_{i=1}^{n} x_i y_i.
\]

For any column vector \(x\) we have

\[
\langle x, x \rangle = x^T x = \sum_{i=1}^{n} x_i^2
\]

and hence \(\langle x, x \rangle \geq 0\). The positive square root is denoted by
Appendix B: Raw Data for Exercise

\[ \|x\| = \sqrt{\langle x, x \rangle} = \sqrt{\sum_{i=1}^{n} x_i^2} \]

and is called the (Euclidean) norm of \( x \).

As a particular case of matrix multiplication, for an \( m \times n \) matrix \( X \) and an \( n \)-dimensional column vector \( x \) we define an \( m \)-dimensional column vector \( Xx \). Moreover, it is shown by calculation that for \( m \)-dimensional column vector \( y \) we have

\[ \langle y, Xx \rangle = \langle X^Ty, x \rangle, \]

where the left-hand side is the inner product of \( m \)-dimensional column vectors and the right-hand side is the one of \( n \)-dimensional column vectors.

A.5 Metric

Measuring difference between two objects is of fundamental importance. For that purpose a metric or a distance function is the most basic tool in various aspects. Let \( X \) be a set of objects or more generally elements. A metric on \( X \) is a function \( d : X \times X \to \mathbb{R} \) satisfying the following conditions:

(M1) \( d(x, x) = 0 \) and \( d(x, y) = 0 \) if and only if \( x = y \);

(M2) \( d(x, y) = d(y, x) \);

(M3) (triangle inequality) \( d(x, y) \leq d(x, z) + d(z, y) \).

Condition (M1) means that the metric separates two distinct elements in \( X \). A function \( d(x, y) \) satisfying (M1)–(M3) except the second half of condition (M1) is called pseudo-metric, where two distinct elements are not necessarily separated by means of \( d(x, y) \).

Let \( \mathbb{R}^n \) be the \( n \)-dimensional coordinate space, where every point is associated with an \( n \)-tuple of real numbers, i.e., the coordinate \( (x_1, \ldots, x_n) \). For two points in the \( n \)-dimensional coordinate space, which are identified with column vectors \( x \) and \( y \) as in (A.5), we set

\[ d(x, y) = \|x - y\|, \quad x, y \in \mathbb{R}^n. \] (A.7)

As is easily verified, the above function \( d \) satisfies conditions (M1)–(M3). Thus, \( d(x, y) \) defined by (A.7) is a metric on \( \mathbb{R}^n \) and is called the Euclidean metric or Euclidean distance. Using coordinate the Euclidean metric is given by

\[ d(x, y) = \left( \sum_{i=1}^{n} (x_i - y_i)^2 \right)^{1/2}, \quad x = [x_1, x_2, \ldots, x_n]^T, \quad y = [y_1, y_2, \ldots, y_n]^T. \]

For \( n = 2 \) or \( n = 3 \) the above relation reflects the Pythagorean theorem.

The Euclidean metric admits a one-parameter deformation. For \( 1 \leq p < \infty \) we define

\[ d_p(x, y) = \left( \sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p}, \]

and for \( p = \infty \) we set

\[ d_\infty(x, y) = \max\{|x_i - y_i|; 1 \leq i \leq n\}, \]

where \( x = [x_1, x_2, \ldots, x_n]^T \) and \( y = [y_1, y_2, \ldots, y_n]^T \) are points in \( \mathbb{R}^n \). It is proved that \( d_p(x, y) \) is a metric on \( \mathbb{R}^n \) for any \( 1 \leq p \leq \infty \). The Euclidean metric is just the case of \( p = 2 \). The metric \( d_1 \) is also referred to as the block distance.

Finally, we mention another distance often appearing in data analysis. Let \( W \) be a set of \( p \) letters, say, \( a, b, \ldots, c \). A concatenation of letters in \( W \) is called a word. Let \( W^n \) be the set of words of length \( n \). A word in \( W^n \) is of the form:

\[ x = x_1x_2\ldots x_n, \quad x_1, \ldots, x_n \in W. \]

Given two words, the number of steps to transform one to another by changing letters is of interest. The number of such steps is given by

\[ d(x, y) = ||\{1 \leq i \leq n; x_i \neq y_i\}||, \quad x = x_1x_2\ldots x_n, \quad y = y_1y_2\ldots y_n. \]

It is easy to see that the above function \( d \) is a metric on \( W^n \), and is called the Hamming distance.

Appendix B: Raw Data for Exercise

The following table contains the raw data used for illustration of statistical analysis.
| No. | height (cm) | weight (kg) | age (year) |
|-----|-------------|-------------|------------|
| 1   | 178         | 100         | 33         |
| 2   | 185         | 90          | 22         |
| 3   | 190         | 90          | 29         |
| 4   | 175         | 79          | 27         |
| 5   | 185         | 81          | 29         |
| 6   | 196         | 106         | 30         |
| 7   | 188         | 100         | 31         |
| 8   | 186         | 87          | 23         |
| 9   | 188         | 84          | 19         |
| 10  | 182         | 77          | 31         |
| 11  | 180         | 82          | 19         |
| 12  | 176         | 82          | 30         |
| 13  | 186         | 85          | 25         |
| 14  | 178         | 88          | 31         |
| 15  | 177         | 77          | 26         |
| 16  | 180         | 77          | 35         |
| 17  | 187         | 108         | 25         |
| 18  | 175         | 68          | 26         |
| 19  | 177         | 78          | 39         |
| 20  | 193         | 105         | 25         |
| 21  | 186         | 96          | 24         |
| 22  | 174         | 78          | 32         |
| 23  | 178         | 85          | 26         |
| 24  | 177         | 81          | 18         |
| 25  | 175         | 81          | 27         |
| 26  | 182         | 86          | 18         |
| 27  | 183         | 80          | 19         |
| 28  | 179         | 93          | 27         |
| 29  | 173         | 74          | 24         |
| 30  | 182         | 84          | 24         |
| 31  | 180         | 89          | 27         |
| 32  | 178         | 80          | 27         |
| 33  | 178         | 85          | 23         |
| 34  | 168         | 69          | 24         |
| 35  | 189         | 90          | 18         |
| 36  | 174         | 74          | 24         |
| 37  | 172         | 72          | 20         |
| 38  | 184         | 74          | 25         |
| 39  | 185         | 88          | 28         |
| 40  | 173         | 72          | 29         |
| 41  | 172         | 75          | 18         |

| No. | height (cm) | weight (kg) | age (year) |
|-----|-------------|-------------|------------|
| 42  | 183         | 78          | 25         |
| 43  | 190         | 84          | 19         |
| 44  | 183         | 83          | 23         |
| 45  | 180         | 80          | 36         |
| 46  | 175         | 77          | 28         |
| 47  | 172         | 96          | 20         |
| 48  | 182         | 90          | 29         |
| 49  | 185         | 92          | 27         |
| 50  | 174         | 79          | 25         |
| 51  | 178         | 82          | 29         |
| 52  | 178         | 83          | 30         |
| 53  | 180         | 73          | 22         |
| 54  | 178         | 76          | 23         |
| 55  | 180         | 93          | 25         |
| 56  | 180         | 74          | 23         |
| 57  | 181         | 85          | 25         |
| 58  | 173         | 75          | 24         |
| 59  | 177         | 86          | 22         |
| 60  | 181         | 75          | 25         |
| 61  | 173         | 73          | 39         |
| 62  | 180         | 79          | 23         |
| 63  | 183         | 85          | 24         |
| 64  | 180         | 75          | 30         |
| 65  | 175         | 75          | 37         |
| 66  | 185         | 85          | 21         |
| 67  | 185         | 86          | 24         |
| 68  | 175         | 73          | 25         |
| 69  | 178         | 73          | 18         |
| 70  | 178         | 85          | 32         |
| 71  | 172         | 70          | 30         |
| 72  | 178         | 87          | 22         |
| 73  | 177         | 85          | 35         |
| 74  | 174         | 82          | 22         |
| 75  | 171         | 75          | 26         |
| 76  | 176         | 78          | 24         |
| 77  | 185         | 85          | 33         |
| 78  | 179         | 80          | 30         |
| 79  | 180         | 96          | 23         |
| 80  | 175         | 79          | 30         |
| 81  | 185         | 82          | 23         |
| 82  | 180         | 93          | 24         |
| 83  | 184         | 85          | 17         |