Probability Theory and Statistics 1

Drs. L.P. Seelen & Dr. G.J.M. Marée

0 Introduction

'Probability Theory and Statistics 1' is the first course in a series of three. Probability theory ánd statistics? To lay people, probability theory is often considered to be part of statistics. Many authors of text books give their books a title like 'Managerial Statistics', but an important part of their books is actually about probability theory. As so often, it is mainly a matter of definition. Here, we will use the following simple definitions: while Probability theory is the study of random phenomena, Statistics is dedicated to get information from data. And the field of Statistics can be split into two different areas, i.e. descriptive statistics and inferential statistics. The latter is, confusingly, often simply referred to by the term 'statistics'….

Before we can briefly introduce these different areas, we must first pay attention to the very important distinction between population and sample. A **population** is a set of similar items which are of interest for a certain research question. A population can be a group of actually existing objects (e.g. the adult Dutch male population) or a hypothetical and potentially infinite group of objects (for example, the set of all possible outcomes which may occur when throwing a specific die an infinite number of times). Also, measurements which could be performed on each item within a population can be viewed as a population in itself, for example the set of *heights* of all adult Dutch men. A population is generally characterized by one or more parameters. A **parameter** is a descriptive measure of a population (a numerical characteristic that can help in defining a population), for example, "the average height of all the Dutch adult males, or "the fraction of the number of times a roll of a particular die results in six dots, or the parameter μ of a Poisson distribution (see section 3.7.7).

A **sample** (Dutch: steekproef) relates to a subset of a population, e.g. ten men selected from the male Dutch population, or the result of a single roll with a specific die. The elements in the sample are called the observations. We talk about a **random sample** if at any stage in the sampling process there is no preference for any element in the population whatsoever. Just as descriptive measures can be defined for a population, we can define descriptive measures for a sample. However, such a measure is now no longer called a parameter, but a **statistic** (Dutch: steekproefgrootheid). It is essential to be always very much aware when we are dealing with a population and when with a sample. We will introduce later even different notations and symbols for otherwise very similar numerical measures, like for example $\overline{x}$ for the sample mean and μ for the population mean. And for the size of the sample, we will use the small letter n, while the capital N is used to denote the size of the population.

Returning to the different areas, we can distinguish:

- **Descriptive statistics** (Dutch: beschrijvende statistiek). This area is engaged in organising, summarising, and presenting data (from either a sample or the population), by means of figures, charts, tables and/or descriptive measures. Note that this conforms to our definition of statistics: it makes the information which is hidden in the data more readily available to the reader. Descriptive statistics is skipped almost entirely in the book of Bain & Engelhardt; in this reader, we pay some attention to it (in Chapter 1), since not every beginning student is sufficiently competent in this field.

- **Probability theory** (Dutch: kansrekening). Probability theory is the study of random phenomena and as such deals with situations where there is no certainty about the outcome of a particular process (experiment or taking a sample). For example, probability theory can make a statement concerning the probability of drawing three spades one after the other from a complete and well-shuffled deck with 52 playing cards. We can recognise in this example a population (the deck of cards), from which we select without replacement a random sample of size 3. In general, probability theory can only really *calculate* a probability if the population as a whole is well known (either by knowing all the items in the population, or by knowing its distribution as well as the parameters of the population). Probability theory will make up the major part of the course Probability Theory and Statistics 1, and will be continued in Probability Theory and Statistics 2.

- I**nferential statistics** (Dutch: verklarende of inferentiële statistiek), or sometimes plainly called statistics. This involves drawing conclusions (in a mathematically correct way)

concerning (the parameters of) a population, using the observations in one or more random samples. Note that, in a sense, this is exactly the opposite of what happens in probability theory. Now the population is not assumed to be known, and instead of making a statement about the probability of a specific sample to happen, we observe a sample and use it to make a statement about the population. This situation is very common, since it is usually impossible or too expensive to investigate a population as a whole. The concepts and especially the validity of the techniques from the area of inferential statistics can only be properly understood if a thorough understanding of probability theory has been reached. Inferential statistics is addressed in the courses Probability Theory and Statistics 2 and 3.

This reader does not always follow the same sequence of topics as the book of Bain & Engelhardt, Introduction to Probability and Mathematical Statistics, 2-nd edition. In order to be able to use the book next to the reader, an effort has been made in this text to refer to the book of Bain & Engelhardt at every section, theorem and definition (see the abbreviation B&E).

1 Descriptive statistics

Descriptive statistics is concerned with organizing, summarizing and presenting data (from sample or population) by the means of graphs, tables and/or descriptive measures. A great number of techniques are part of descriptive statistics, of which we will only discuss the most important ones in this text.

1.1 Scales of measurement

Statistics is all about extracting information from data. But there are different types of data, and those types also determine the techniques which can be employed in order to extract the information from the data.

A **variable** is a characteristic of a population or a sample, for example, the variable "height" of an adult Dutch male. The height will vary from individual to individual, hence the name *variable*. The **values** of a variable are the possible observations of that variable. If the height is measured in centimetres, then the range of values for the height consists of the interval of, say, 100 cm to 250 cm. The **data** are the actually observed values of a variable in a specific population or in a specific sample of the population. Although people may automatically think of numbers when talking about data, data can also be non-numerical. The variable 'country of birth' has for example as possible values: Netherlands, Germany, Japan, Iran, etc. We use the term **scales of measurement** (Dutch: meetschalen**)** to make a distinction between those different types of data. We will give a classification here which is fairly widely used, although you should not be surprised to find other classifications as well.
Where data are inherently numerical in nature, we talk about **quantitative** data, like the height data of Dutch adult males. If not, then we talk about **qualitative** data. Both types can be divided into different subtypes.

Qualitative scales

The values of data on the qualitative scale are categories, like we have seen already for the variable Country of birth. Although it is possible to assign numbers to each of those categories (for example: 1=Netherlands, 2=Germany, 3=Japan, etc.), that does not turn the data into quantitative data, because it is still useless to talk about the average country of birth by calculating the average of all those assigned numbers in a sample. Note that many different coding systems may exist (for example: 23=Netherlands, 8=Germany, 1034=Japan, etc.), and that also the order in which the countries are listed is very arbitrary. Country of birth is an example of a variable measured on the **nominal** scale. Below, we will see that nominal data allow for only very few techniques from the field of descriptive statistics.

But qualitative data can also be of the **ordinal** subtype, where a natural order *does* exist. In an opinion poll, the respondent may be asked for his opinion about a statement with the possible answers 1=strongly disagree, 2=disagree, 3=neutral, 4=agree and 5=strongly agree. But even here, the coding is arbitrary; we could have chosen as well the coding 11=strongly disagree, 25=disagree, 30=neutral, etc. as long as the order is maintained.

Remark: in practice, ordinal data are often treated as if they were truly quantitative; this is only acceptable in cases where we have good reason to think that the differences between two successive categories always have the same 'magnitude', whatever that may mean.

Quantitative scales

A quantitative scale can either be discrete or continuous. The data are said to be **discrete** if the possible values for the variable consists of a finite set, or an infinitely large but countable set. Very often, the data will consist of integers only. For example, the number of correct answers on a particular test with 20 questions (set of possible values is finite: {0, 1, 2, ..., 20}), or the number of cars driving through a tunnel per year (infinitely large, countable set: {0, 1, 2, etc.}).

If a quantitative variable can take on any value, usually within a certain range, we say that it is measured on a **continuous** scale. For example, the time that elapses between two successive collisions of two atoms, measured with an infinitely great accuracy. The height of adult Dutch males can also be

considered as a continuous variable. In practice, however, height measurements are performed with limited accuracy, for example in millimetres. Strictly speaking, rounding of the true heights results in discrete data, but generally we will still regard those as being continuous.

Apart from the distinction discrete versus continuous, quantitative data can also be thought of as measured on an interval or on a ratio scale. Data are measured on an **interval** scale if *differences* between values are consistent and meaningful, but *ratios* are not. For example, the difference in temperature between 12˚ C and 17˚ C is the same as the difference between 30˚ C and 35˚ C. Note that *ordinal* data are *not* interval data, exactly because it is not possible to say that the (size of the) difference between 'neutral' and 'agree' is the same as between 'agree' and 'strongly agree'.
When data are measured on a **ratio** scale, then not only differences between values are meaningful, but also the ratios of two values. A variable length is a ratio variable since it is meaningful to say that 12 cm is three times as long as 4 cm. But we cannot really say that a temperature of 12˚ C is three times as warm as 4˚ C, so temperature in degrees Celsius is not a ratio scale. However, temperature in degrees Kelvin is; note that a ratio scale has a meaningful zero value, which an interval scale does not necessarily have.
However, this distinction (interval vs ratio) is in itself not relevant for most statistical techniques; some books therefore only use the term interval scale, even when the variable itself is measured on a ratio scale.

1.2 Techniques for a single variable

Graphs and tables are intended to present data in such a way that the information contained within the data is communicated as clearly as possible to the reader. That should always be the decisive factor when designing graphs or tables, so knowing the targeted group of readers is important. Apart from graphs and tables, we can also try to summarize certain characteristics of the data in so-called measures. The average height of all Dutch males is an example of such a measure. We are primarily interested in two types of measures: measures of location and of spread. A **measure of location** (or central tendency; Dutch: locatiemaat of centrummaat) is intended to summarize data by a certain 'typical value', such as the mode, median and mean which will be discussed below. A **measure of spread** (or dispersion; Dutch: spreidingsmaat) tells us something about the amount of variability within the data set. We will discuss here very briefly the most important techniques for each of the different measuring scales.

1.2.1 Nominal data

For nominal data, only relatively few techniques are available; we will use the following example to discuss the most important ones. There is no way in which we can define some kind of 'average' when dealing with nominal data. The only measure of location is the **mode** (Dutch: modus). The mode is defined as simply the value which occurs most frequently (or the category with the largest number of observations).

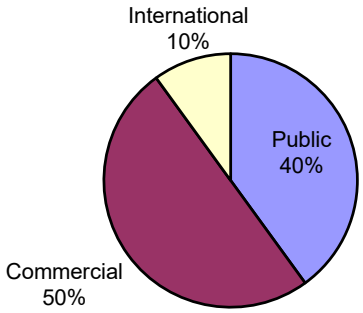
Example 1.1

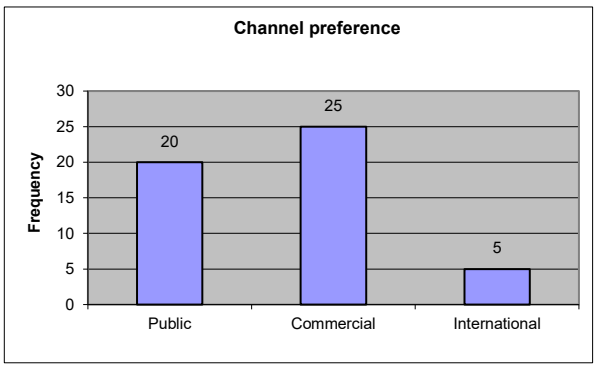
A sample of 50 people were asked whether they prefer to watch either public tv-channels (P), commercial channels (C), or international channels (I). The raw data of this sample are as follows: {C, C, P, P, C, C, P, C, C, C, P, P, P, C, P, P, P, C, C, C, C, C, P, I, C, P, C, C, I, P, C, C, C, P, P, I, C, P, I, P, C, P, C, P, C, P, C, C, P, C, I}.

We can use the raw data to create a **frequency table** (the column 'relative frequency' is optional):

Channel	Frequency	Relative frequency
Public	20	0.40
Commercial	25	0.50
International	5	0.10

The mode is here 'Commercial', since that is the most frequent item. Graphically, we can create either a pie chart (on the left; Dutch: cirkeldiagram) or a bar chart (on the right; Dutch: staafdiagram).





1.2.2 Ordinal data

The same graphic techniques can be applied as for nominal data, although a bar chart is now more appropriate than a pie chart. With a frequency table, there is now the additional option of including a column with the 'cumulative relative frequency'. As a measure of location, it is now also possible to define the **median** in a meaningful way as the middle-ranked item, or that value which separates the lower half of the data from the higher half.

Example 1.2

Say $n = 16$ students are asked for their opinion about the statement 'Statistics is fun!', where the choice is between strongly disagree, disagree, neutral, agree or strongly agree. This might result in the following frequency table:

	Frequency	Rel. Freq.	Cum. Rel. Freq.
strongly disagree	3	0.1875	0.1875
disagree	1	0.0625	0.25
neutral	3	0.1875	0.4375
agree	4	0.25	0.6875
strongly agree	5	0.3125	1

For example, the cumulative relative frequency for the category 'neutral' can be found by summing all the relative frequencies for the categories 'Strongly disagree' to 'neutral'.
The mode is now 'strongly agree', while the median is 'agree' (it is the first category where the cumulative relative frequency exceeds 0.5) ◄

1.2.3 Quantitative data

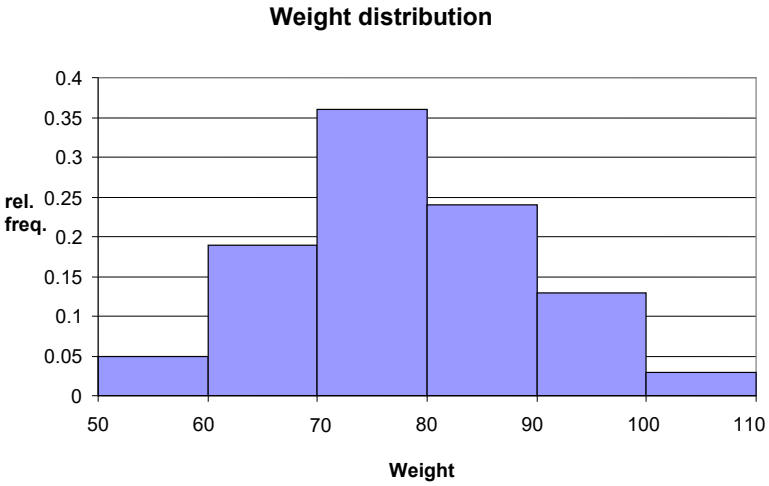
Charts and tables

With discrete data, we can apply in principle the same charts and tables we already know for ordinal data. When dealing with continuous data, but also with discrete data with many different possible values, we encounter another aspect: there is now no longer an unambiguous way to divide the data into categories. Suppose for example that we have weight data of $N = 200$ people, ranging from 52.1 kg to 107.7 kg. In order to determine the categories (which we now call **classes),** we must keep in mind that the goal is to obtain a table or graph which can be easily interpreted. As no clearly defined method exists for doing so, we can only give some guidelines. The limits for each class should be as simple as possible, for example: class 1 from 50 to 60 kg, class 2 from 60 to 70 kg, etc. In general, it is also recommended to use classes with equal widths (like the widths of 10 kg above; the only exception might be when that would lead to classes that contain no or almost no observations). Furthermore, it is common to choose a larger *number* of classes when the data set itself is large. At the same time this number should not become too big, in order to prevent poor readability. As a result, we could obtain the following frequency table, which seems to represent reasonable choices:

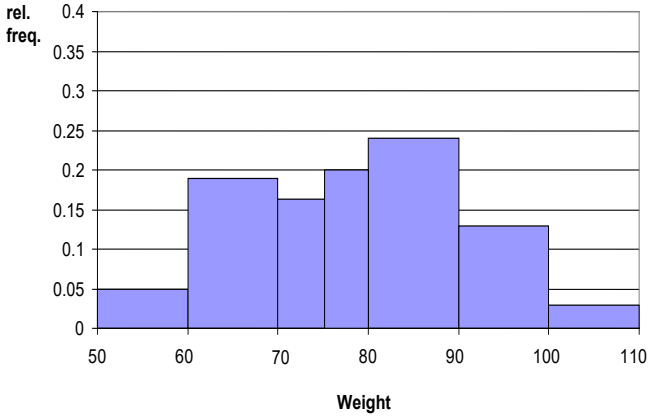
i	Class limits	m_i	f_i	rf_i	crf_i
1	50 - 60	55	10	0.05	0.05
2	60 - 70	65	38	0.19	0.24
3	70 - 80	75	72	0.36	0.60
4	80 - 90	85	48	0.24	0.84
5	90 - 100	95	26	0.13	0.97
6	100 - 110	105	6	0.03	1.00
total			200	1.00	

The column m_i shows the midpoint of class i ; although it is not essential at all, it can be convenient when we need to calculate some measures later on. The other columns should be clear. In the example above, we have assumed that the weight data were measured with an infinitely great precision. In that case no person will weigh *exactly* 60 kg; in other words, it will always be clear to which class a particular observation belongs. In practice, however, we usually are dealing with numbers which have been rounded, for example, to the nearest integer. So someone with a measured weight of 60 kg could actually weigh anywhere between 59.5 kg and 60.5 kg. This will have consequences for the class limits as shown in the table: the first class would then become the class of 50 - 59 kg, the second from 60 to 69, etc. Such a classification would be exactly the same as the following: the first class from 49.5 to 59.5 kg, the second from 59.5 to 69.5, etc. The class midpoints m_i are then 54.5, 64.5, etc.

Whenever we are dealing with quantitative data, we no longer talk about bar charts (which have gaps between the columns) but about **histograms** (without gaps, like seen in the figure below).



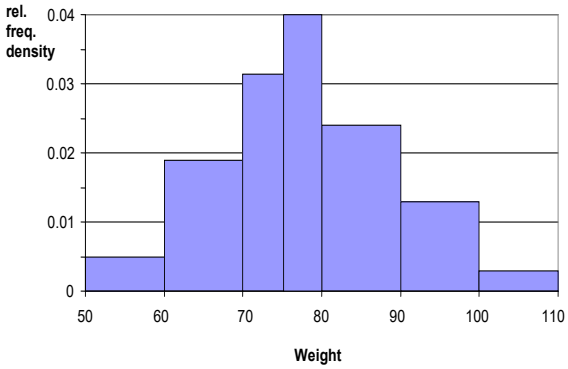
When we draw a histogram, we can choose to display on the vertical axis either the frequency or the relative frequency. This choice affects only the numbers along the axis; the shape of the histogram remains unchanged. But what happens if we split a single column into two more narrow columns? For example, if we split the class of 70 to 80 kg in a class of 70 to 75 kg, and another class from 75 to 80 kg? Since the frequency in the old class (72) is now divided over two classes, the frequencies for each of the two new classes will necessarily become lower (say e.g. 32 and 40), and thereby also changing greatly the shape of the histogram (see figure to the right). This is unsatisfactory, because the shape of the histogram no longer seems intuitively easy to understand.



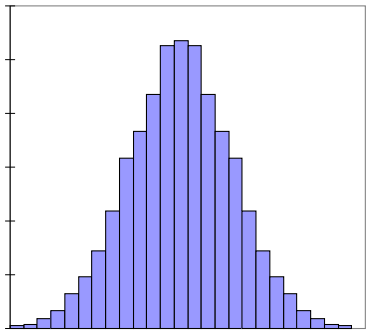
However, this can be remedied by displaying on the vertical axis the so-called **(relative) frequency density**, instead of the (relative) frequencies. These densities are simply found by division of the (relative) frequency

by the class width. Thus, the reduced frequency which resulted from the split of a class is now compensated by the reduced class width.
Note that the surface area of each column in the figure below is equal to the relative frequency of the class, which means that the total area of the columns adds up to exactly 1! So, whenever we are dealing with a frequency table with unequal class width, we should use the densities on the vertical axis instead of the frequencies themselves.

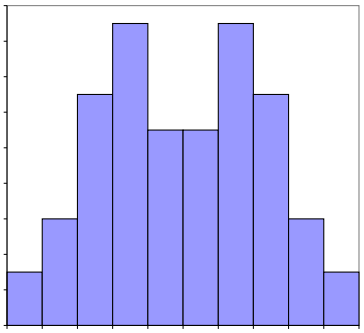
i	Class limits	f_i	rf_i	$density_i$
1	50 - 60	10	0.05	0.005
2	60 - 70	38	0.19	0.019
3	70 - 75	32	0.16	0.032
4	75 - 80	40	0.20	0.040
5	80 - 90	48	0.24	0.024
6	90 - 100	26	0.13	0.013
7	100 - 110	6	0.03	0.003
total		200	1.00	



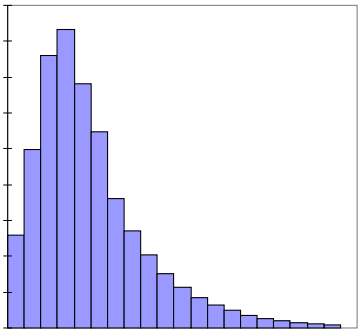
Just looking at the shape of a histogram tells us some interesting things about the data set. E.g.:



Bell shape: symmetric and uni-modal

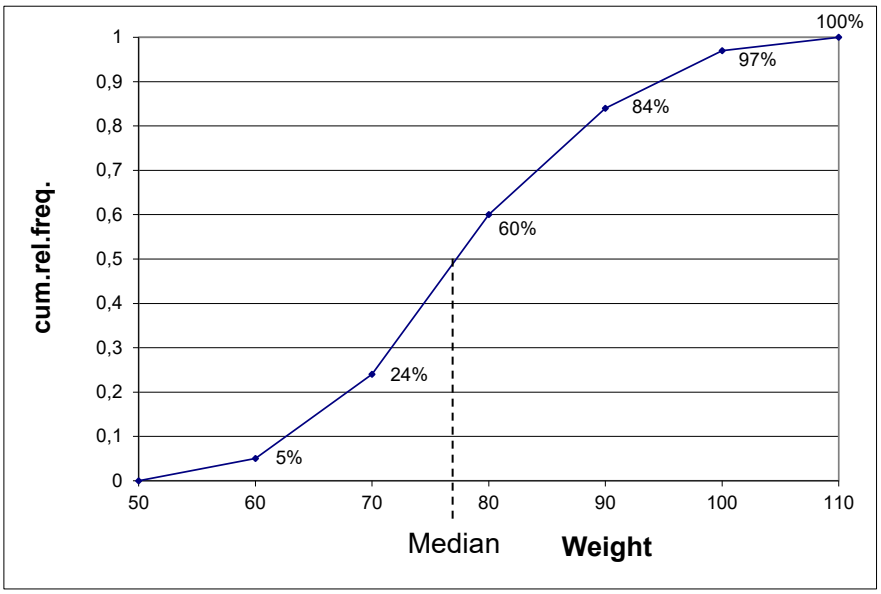


Symmetric, bi-modal



Skewed to the right, Or positively skewed

A graph that represents the cumulative frequency or cumulative relative frequency for each class is called an **ogive** or **cumulative relative frequency polygon** (Dutch: ogief). It is constructed by plotting points whose x-coordinates are the upper class limits and whose y-coordinates are the cumulative frequencies or cumulative relative frequencies. The ogive can be used, for example, to estimate the median, by dropping a vertical line from the ogive at a cumulative relative frequency 0.50:



Measures of location

In contrast to nominal and ordinal data, quantitative data allow for different types of calculations. The most common measure of location is the (arithmetic) **mean**: the sum of all observations divided by the number of observations.

Sample mean: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and Population mean: $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$

(Dutch: steekproefgemiddelde en populatiegemiddelde).

Note the very important difference in the symbols we use to denote the mean; in general, we will use Greek letters for measures of a population, and Latin letters for a sample (also remember that n stands for the size of a sample taken from a population, and N for the size of the total population).

Remark. Other means exist apart from the arithmetic mean. For example, in situations where we need to determine the average of multiplication factors (such as in the calculation of the average return on an investment over a number of years), we use the geometric mean. And the average of many ratios can better be measured by the harmonic mean. Within this text, however, we do not pay any attention to those other means.

Of course we can still use the previously mentioned **median,** which is the number separating the lower half of the data set from the upper half. We may also use **percentiles**: a percentile of a data set is one of the 99 points which splits the ordered data set into 100 parts, each with an equal number of observations. The k^{th} percentile is a number that separates the k% smallest data from the (100−k)% largest data. The 95th percentile, for example, is a number such that (at most) 95% of the observations are less than it, and (at most) 5% are greater than it. The 50th percentile is the same as the median. And the 25th , 50th and 75th percentiles are also called the 1st, 2nd, and 3rd **quartile** respectively**.** In practice, a percentile is often not entirely determined by the above description, and we will have to use some interpolation method. However, there are several ways of doing this; here, we will use basic linear interpolation. The larger the data set, the less difference there will be between the different methods

Example 1.3

The frequency table on page 6 shows that 60% of all individuals in the data set had a weight of less than 80 kg, and 84% had a weight of less than 90 kg. The 75th percentile (= 3rd quartile) of the data set must be somewhere within the class of 80 to 90 kg. Linear interpolation gives a value of 86.25 kg (= 80 + (75–60)/(84–60)*(90–80). Similarly, the 1st quartile will be around 70.28 kg. ◄

Example 1.4

What is the median of the following 6 numbers: 1, 2, 3, 4, 5, and 6? According to the way the median has been introduced here, any number between 3 and 4 (including the numbers 3 and 4) could be chosen as the median. However, in this example it makes sense to take the middle value of the range 3 to 4, so we could say that 3.5 is the median. ◄

While there are no more than 99 percentiles, **quantiles** generalise this idea further. The p-quantile (with p between 0 and 1) is the value such that a proportion p of all observations is smaller than or equal to that value. Thus, the 95th percentile is the same as the 0.95-quantile and the median is the 0.5-quantile.

Measures of spread

Measures of spread (Dutch: spreidingsmaten) are used to describe the variability in a sample or population. The **range** (Dutch: spreidingsbreedte) is defined as the difference between the largest and the smallest value in the data set. But a much more common measure is the **variance** (Dutch: variantie). There is a subtle but important difference between the definition of the sample variance, and the population variance:

Sample variance: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$ Population variance: $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$

The reason for the remarkable difference in the denominators will only become really clear during the course Probability Theory and Statistics 2; for now, it is sufficient to say that the sample variance as defined in this way provides the best possible estimate of the population variance (remember that we are usually not able to study the population as a whole, so the only way to estimate the population variance is taking a sample and use it to determine the sample variance).
In both formulas, we take for each observation the difference between the value and the mean of all values. Subsequently, these differences are squared and summed. It will be immediately clear that the variance can never be negative. And the variance can only ever be equal to 0, if all the observations are exactly the same. The more an observation deviates from the mean, the more it contributes to the size of the variance.

There are alternative formulations which often require fewer calculations, because it is not necessary for each observation to determine and square the difference in relation to the mean:

Sample variance: $s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right)$ Population variance: $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \mu^2$

We show below that the two formulas for the sample variance are equal to each other:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i^2 - 2x_i\overline{x} + \overline{x}^2)$$
$$= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} 2x_i\overline{x} + \sum_{i=1}^{n}\overline{x}^2\right)$$
$$= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - 2\overline{x}\sum_{i=1}^{n} x_i + n\overline{x}^2\right) \qquad \text{(Recall that: } \sum_{i=1}^{n} x_i = n\overline{x}\text{)}$$
$$= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - 2n\overline{x}^2 + n\overline{x}^2\right) = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right)$$ ◄

Example 1.5

Calculate the sample variance for the next sample of 5 observed lengths (in cm):

6, 10, 3, 4, 6

We first determine the sample mean: $\overline{x} =$ (6+10+3+4+6)/5 = 5.8 (cm),

According to the first formula for the sample variance, we obtain:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{5-1}\left((6-5.8)^2 + (10-5.8)^2 + (3-5.8)^2 + (4-5.8)^2 + (6-5.8)^2\right)$$
$$= \frac{1}{4}\left(0.2^2 + 4.2^2 + 2.8^2 + 1.8^2 + 0.2^2\right) = 7.2 \text{ (cm}^2\text{)}$$

According to the second formula:

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right) = \frac{1}{4}(6^2 + 10^2 + 3^2 + 4^2 + 6^2 - 5 \cdot 5.8^2) = 7.2 \text{ (cm}^2\text{)}$$

Note that the unit of measurement of the variance here is equal to cm x cm = cm^2. If the lengths would have been measured in mm instead, then the variance would be 100 times as large, so 720 (mm^2). ◄

The above example illustrates that the variance is very sensitive to the unit of measurement used. Because the units are squared, the magnitude of the variance is a quantity that is very hard to interpret. But if we take the square root of the variance, a quantity results that it is easier to interpret, the **standard deviation**:

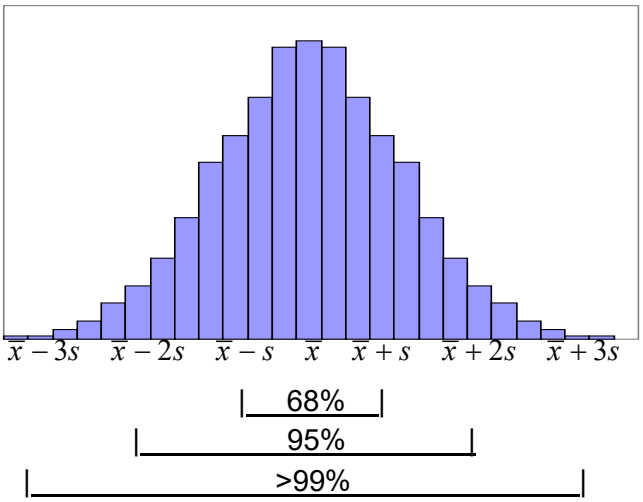
Sample standard deviation: $s = \sqrt{s^2}$ Population standard deviation: $\sigma = \sqrt{\sigma^2}$

Example 1.6

The sample standard deviation for the length data in Example 1.5 is 2.683 (cm) or 26.83 mm. ◄

Thus, the unit for the standard deviation is the same as the unit of measurement of the data itself. That makes some statements possible. When a histogram is approximately bell-shaped, we can, by approximation, make the following statements:

About 68% of all observations lie within the interval $(\bar{x} - s,\ \bar{x} + s)$ (or $(\mu - \sigma,\ \mu + \sigma)$)
About 95% of all observations lie within the interval $(\bar{x} - 2s,\ \bar{x} + 2s)$ (or $(\mu - 2\sigma,\ \mu + 2\sigma)$)
More than 99% of all observations lie within the interval $(\bar{x} - 3s,\ \bar{x} + 3s)$ (or $(\mu - 3\sigma,\ \mu + 3\sigma)$)



Later it will become clear that the above percentages can be derived directly from the standard normal distribution. For distributions without bell shapes, we cannot use the above, but we can still make some statements using Theorem 3.6 (see later; Inequality of Chebyshev).

There also exists a unitless measure of spread, which is however not widely used in practice, i.e. the **variation coefficient**:

Sample variation coefficient: $cv = s / \bar{x}$ Population variation coefficient: $cv = \sigma / \mu$

The last measure of dispersion we mention briefly is the **interquartile range, IQR** (Dutch: interkwartielafstand) which is defined as the difference between the 3rd and the 1st quartile.

Grouped data

If the original (raw) data are not available, but only a frequency table like the one on page 6, then we cannot use the formulas for the mean and the variance as discussed above. However, we can still try to estimate those measures. When we only know that, say, 10 observations fall within the class 50 to 60 kg, but we do not know the exact observations, then the best bet is probably to assume that the average value of those 10 observations is equal to the class midpoint (in this case 55 kg). That leads to the formulas below:

Sample mean: $\bar{x} = \frac{1}{n}\sum_{i=1}^{k} f_i m_i$

Sample variance: $s^2 = \frac{1}{n-1}\sum_{i=1}^{k} f_i (m_i - \bar{x})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{k} f_i m_i^2 - n\bar{x}^2\right)$

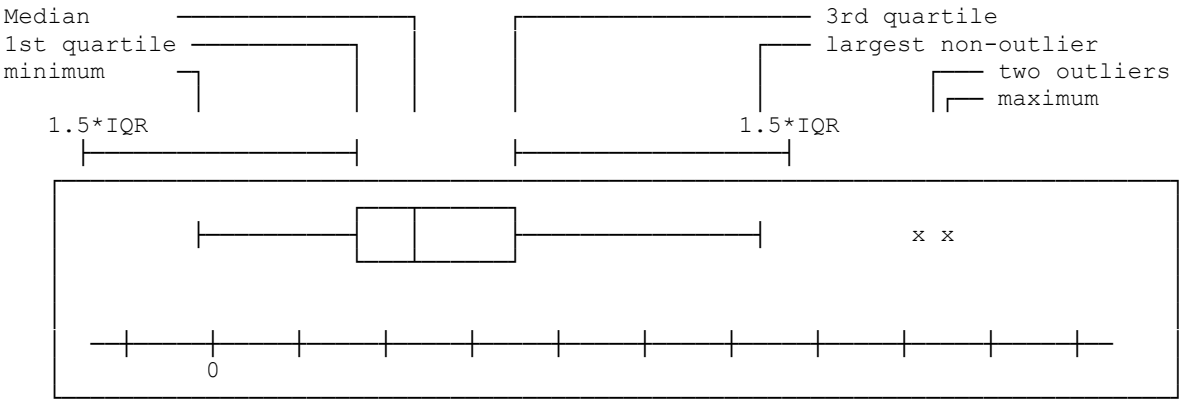
Outliers

In quantitative data sets, it is advisable to check whether some observations are very different in value from the rest of the data. Those observations are called **outliers** or extreme values (Dutch: uitschieters, uitbijters), and are often worth a closer look. Outliers can occur by chance in any sample, but they can also indicate either a measurement error or a population which is very skewed in one direction. In the first case, it may be better to discard them or use special techniques that are robust to outliers, while in the latter case they indicate one should be very cautious in using tools that assume a normal distribution (like linear regression). A huge number of different criteria have been designed to determine if an observation merits this special attention. We mention here only one of those: an outlier is any data point more than 1.5 times the interquartile range below the first quartile or above the third quartile.

Box plot

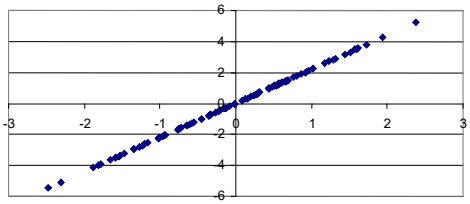
A **box plot** is a widely used graphical representation of a data set, which gives at a glance a rather good impression of the distribution of the data. One axis (below, we chose the x –axis) is the data axis. A rectangle (box) is drawn from the 1st quartile to the 3rd quartile, with an extra line at the median. Then two lines are drawn (the whiskers): the first one runs from the smallest observation which is not an outlier to the left side of the box, and the second one runs from the right side of the box to the greatest observation which is not an outlier. Outliers are then indicated with a dot or star.

The figure below is taken over from Wikipedia:

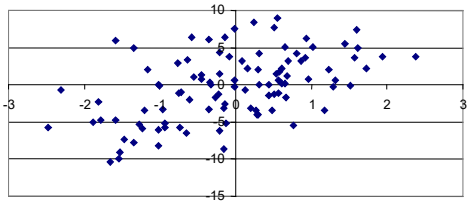


1.3 Techniques for the relation between two variables

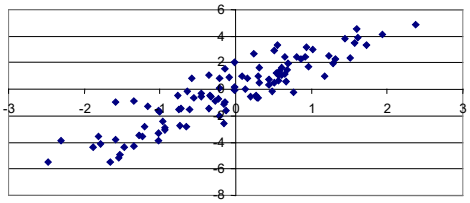
Often, we will be interested in the relationship between two (or more) variables, and again we need to differentiate between the quantitative and the qualitative measurement scales. However, no further techniques for nominal and ordinal variables will be discussed here, but we will focus exclusively on the relationship between two quantitative variables. We assume that the two variables represent different characteristics of the same items, like for example weight and height data for Dutch males; so the data consist of pairs of observations (x_i, y_i). The most widely used graphical representation is the so-called **scatter plot** (Dutch: spreidingsdiagram, puntenwolk, correlatiediagram). The x- variable is plotted on the horizontal axis, and the y variable on the vertical axis. Each pair is displayed as a single dot in this axis-system. When we do this for all pairs, in the end the scatter plot gives an indication of the existence of a possible relationship the two variables.



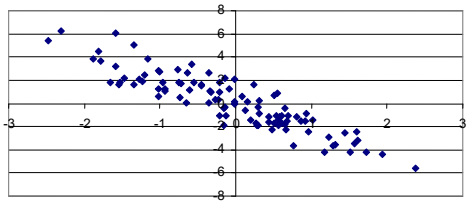
r=1.0, perfect positive relationship



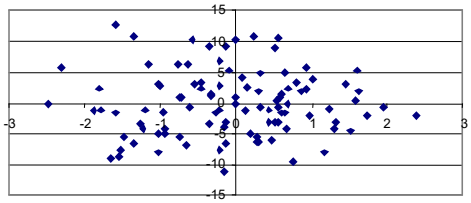
r=0.5, positive relationship



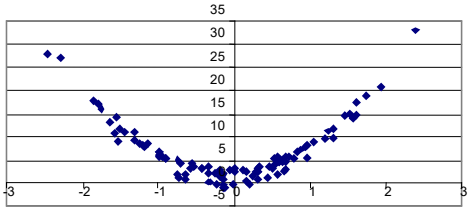
r=0.9, strong positive relationship



r=–0.9, strong negative relationship



r=0.0, no relationship



r=0.0, nonlinear relationship

In each of the above examples, a numerical measure r is mentioned, which denotes the sample correlation coefficient. In order to calculate the correlation coefficient, we must first discuss the covariance. The covariance says something about the linear relationship between two variables. We distinguish again between sample and population:

Sample covariance: $\mathrm{COV}(x,y) = s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$

Population covariance: $\mathrm{Cov}(x,y) = \sigma_{xy} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)$

If relatively large values for x_i (that is, values greater than $\overline{x}$) are often paired with relatively large values for y_i, (and so also vice versa: relatively small values for x_i often coincide with relatively small values for y_i), then in the above formula most of terms $(x_i - \overline{x})(y_i - \overline{y})$ will be positive, so the covariance will be positive as well. However, are relatively large values for x_i often paired with relatively small values for y_i (and vice versa), then most terms $(x_i - \overline{x})(y_i - \overline{y})$ will be negative, so that the covariance will also be negative.

So although the sign of the covariance provides us already with important information, the size of the covariance as a measure of linear relationship is not suitable for interpretation, since it depends directly on the units of measurements used for both variables. So if, for example, we measure the weight and height in a sample of Dutch males in metres and kilograms, the covariance would be a factor 100,000 smaller as compared to the case when the units of measurement were chosen to be

centimetres and grams instead. But by dividing the covariance by the product of the two standard deviations we obtain a measure, called the correlation coefficient, which is unitless and which will not change if we choose other units of measurements for the variables:

Sample correlation coefficient: $r_{xy} = \frac{s_{xy}}{s_x s_y}$

Population correlation coefficient: $\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$

It can be shown (as we will do so for the population coefficient of correlation in the course Probability Theory and Statistics 2), that the above formulas will always result in a value between -1 and +1.

1.4 Problems

1.1 Classify each of the variables below in one of the following categories:
Quantitative-discrete, quantitative-continuous, qualitative-ordinal of qualitative-nominal.

a number of children in a household
b weekly closing price of gold
c size of a soda (small, medium, large)
d amount of oil that was imported monthly in the Netherlands
e mark that was obtained for an oral exam
f year of birth of a person
g civil service salary scales (5 through 16)
h country of origin of a product
i start number of a participant in the 'Elfstedentocht' (scating race along 11 Frisian cities)

1.2 An energy supplier investigates the appreciation of the monthly magazine that is sent to its customers. A sample of 1000 customers leads to the following results:

Appreciation of magazine	Count
very uninteresting	70
uninteresting	215
neutral	375
interesting	210
very interesting	130
total	1.000

a Add to this table a column with the cumulative relative frequencies.
b Find the mode, the median and the mean.

1.3 A producer of bicycle tyres keeps record of the number of defective bicycle tyres per day, during a period of 25 days. The results (i.e. the number of defective bicycle tyres per day) are as follows:

7 9 10 11 12 13 14 14 15 15 16 17 17
18 18 19 19 20 21 21 24 24 26 31 36

a Make a frequency table. (choose the number of categories yourself, at least 5, at most 8, choose 'neat' class bounds)
b Draw a histogram with relative frequencies. Characterize the skewness of the distribution.
c What is the relation between the surface area of each column in the histogram and the relative frequencies?
d Draw the ogive.
e Find proper values of the three quartiles. Are there any outliers?
f Merge the three highest classes, and draw a new histogram. Show that it is now better to put the frequency density on the vertical axis, instead of the relative frequency.

1.4 For a distribution that is skewed to the left, is the median bigger than the mean, or smaller, or equal to the (arithmetic) mean?

1.5 Comparing the mean with the median, which of the two measures is more sensitive to outliers or errors in when entering the data (for example when a typed value is 100 times as big as the correct value)? Which measure makes more efficient use of the available data?

1.6 The number of filled teeth in the mouth of 12 randomly selected Dutch people has the following values:

3, 0, 11, 6, 6, 3, 8, 1, 0, 7, 2, 13

Calculate the average number of filled teeth, and calculate the variance using two different formulas. Also calculate the standard deviation and the range.

1.7 Within an industry in the Netherlands 7 different companies are active. These companies had the following turnover in 2010 (in millions of euros):

12 39 24 92 32 6 51

Calculate the average of the 7 companies and the standard deviation. Do you calculate σ or *s* in this case?

1.8 Show that the sum $\sum_{i=1}^{n}(x_i - a)^2$ achieves a minimum for $a = \bar{x}$. Use this information to explain that it is logical that in the formula for the sample variance $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ the sum is divided by a number that is less than *n*. (Hint: we would actually like to measure the deviation of each observation relative to the *population mean*, but this mean is usually unknown).

1.9 A sample of 76 consultants resulted in the following table:

Income in euro's	Count
25.000 to 40.000	9
40.000 to 55.000	28
55.000 to 70.000	24
70.000 to 85.000	9
85.000 to 120.000	6
Total	76

Calculate (or approximate)

a the sample mean
b the mode / modal class
c the three quartiles
d the sample standard deviation

1.10 Find the mean, median, mode, standard deviation, and coefficient of variation of the sample that is summarized in the following frequency table.

Observed value	4	5	6	7	8	9	10
Frequency	4	15	14	9	5	2	1

1.11 Show that the sample covariance $\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ can also be calculated with the formula $\frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\,\bar{y}\right)$.

1.12 Let *x* be any variable observed for each element in a sample. What is the sample covariance between *x* and *x* equal to? And the value of the correlation coefficient of *x* and *x* ?

1.13 Draw a scatterplot for the next paired observations that were observed for a sample.

$\bar{x} = -0.2$

$\bar{y} = 4$

i	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$x_i y_i$
1	-1	6	-0.8	2	-1.6	-6
2	1	3	1.2	-1	-1.2	3
3	-3	8	-2.8	4	-11.2	-24
4	0	2	-0.2	-2	-0.4	0
5	2	1	2.2	-3	-6.6	2
					-21	-25

Calculate the correlation coefficient between *x* and *y*. Is it what you expected?

$r_{xy} = \frac{S_{xy}}{S_x S_y}$ $S_{xy} = \frac{1}{n-1}\left(\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})\right) = \frac{1}{4}(-21) = -5.25$

$= \frac{-5.25}{\sqrt{5}\sqrt{18.5}} = -0.926$ $= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}\right) = \frac{1}{4}(-25 - 5(-0.2)(4))$

$s_x^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right) = \frac{1}{4}((-0.8)^2+0.2^2+(-2.8)^2+(0.2)^2+2.2^2) = 2.7$

$s_y^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n}(y_i-\bar{y})^2\right) = 8.5$

1.14 For a sample of home-owners, the owner's income and the size of the living area are recorded. The relationship is assumed to be linear. A sample gives the following information:

House number	1	2	3	4	5	6	7
Living area (m2)	100	80	150	200	120	110	80
Owner's income (€1000)	50	45	85	110	60	70	70

For the living area the following statistics can be calculated from the given data: a sample mean of 120 m2 and a sample standard deviation of 42.82 m2.

a Calculate the sample mean and the sample standard deviation of the owner's income.

b Draw a scatter plot with living area on the vertical axis. Why do we put living area on the vertical axis?

c Calculate the covariance and the correlation coefficient.

1.15 Besides the arithmetic mean and the geometric mean, there are some other meaningful means. Here we will derive the formula of the so-called harmonic mean of n numbers. Do this first on the basis of the following example: the distance between Amsterdam and Arnhem and is 100 km. The outward journey is travelled at an average speed of 50 km / h, the return at an average speed of 100 km / h. What is the average speed of the combined outward and return trip? (Hint: Determine first the total distance. Divide this distance by the total time.)
Does the outcome change if the distance between Arnhem and Amsterdam is now set equal to c ? Set up a general formula. Extend that formula in case there are *three* distances of c km, with different speeds (v_1, v_2, v_3).

(also in electrical engineering this mean plays a role: the electrical resistance of parallel resistors remains the same if each is replaced by the harmonic mean)

2 Principles of probability

2.1 Introduction

Games of chance, like throwing dice or playing card games, have been played already for many centuries and were a major reason for the development of the branch of mathematics called probability theory. A proper estimation of probabilities enabled better decision making, and thus also lead in general to higher winnings. Of course, people were already at a very early stage aware that, for example, when throwing two dice it is more likely to obtain a total of 7 dots than any other number of dots. But the mathematical underpinning for many results remained rather unclear until the development of a more theoretical basis from the eighteenth century onwards. A very important work was "Théorie analytique des probabilités", published in 1812 by Pierre-Simon Laplace. But still there was no proper general and axiomatic approach. This modern approach to probability theory was formulated only in 1934 by the Russian mathematician Kolmogorov, who published the book "Grundbegriffe der Wahrscheinlichkeitsrechnung" (in the German language).

2.2 Experiments, sample spaces and events

(B&E, page 1-7)
We can distinguish two basic types of experiments: deterministic and random. A chemist who demonstrates that mixing two specific chemical compounds in water leads to a bright blue colour, is performing a **deterministic** experiment: the resulting colour will always be the same (as long as the experiment is performed in the correct way. But probability theory is dealing with **random** experiments where the results depend on chance (Dutch: kansexperiment, stochastisch experiment). Such an experiment can for example result in:

- the answer which a randomly selected person gives to the question which party he will vote for;
- the number of dots when a die is rolled;
- the amount of rainfall during the next month;
- the waiting time in a queue.

When an experiment is performed repeatedly, the individual repetitions are called **trials**. A fixed number of trials can in turn be viewed as a **composed experiment**. The set of all possible outcomes of (a trial of) the experiment, is called the **sample space** (Dutch: uitkomstenruimte) which we will indicate here by the symbol S (in textbooks, the Greek capital Omega is also used frequently). In one random experiment, it is often possible to define the sample space in more than one way.

Example 2.1

Consider the experiment of rolling a single die. When we are interested in the number of dots, then the most straightforward definition of the sample space would be simply the set of elementary outcomes: S ={ 1, 2, 3, 4, 5, 6 }. But if we are only interested whether six dots appear on the die or not, we can also define S as { 6, not 6 }.
When we perform two trials of this experiment, we can define a combined sample space for this composed experiment with 36 different outcomes as follows: S ={ (1,1), (1,2), (1,3), (1,4), (1,5), (1,6), (2,1), …. }. (Here the outcome (1,3), for example, means that the first roll resulted in 1 dot and the second in 3 dots). But in case we are only interested in the sum of dots in both trials, then we can also use the sample space S = { 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12 }. ◄

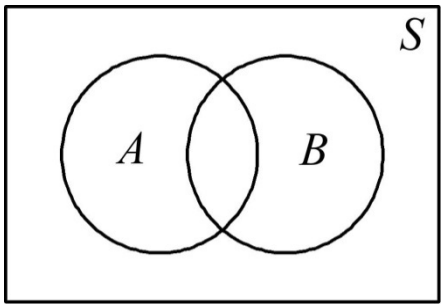
A sample space is called finite if it contains a finite number of different outcomes, such that we can write $S = \{e_1, e_2, \ldots, e_N\}$. The space is called countably infinite if it is possible to order the infinite number of outcomes in such a way that each outcome has a specific position in $S = \{e_1, e_2, \ldots\}$. For example, if an experiment results in a natural number (without being able to put any upper limit to the size of the largest number, then we can define S = { 0, 1, 2, 3, …}. However, S = { 0, 2, 1, 4, 6, 3, 8, 10, 5, …. } is another possible definition for this situation.

In all cases above, we were dealing with **discrete sample spaces**. But if we, for example, perform an experiment by measuring (with infinite accuracy) the time until a radioactive atom falls apart, we are dealing with a **continuous sample space** , for example, $S = \{ t \mid t > 0\}$.

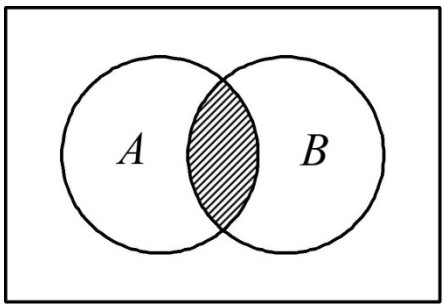
A subset of a sample space is called an **event** (Dutch: gebeurtenis). Examples of events when rolling a die are: 'the number of dots is greater than 3' and 'the number of dots is even', with corresponding subsets of the sample space $\{4, 5, 6\}$ and $\{2, 4, 6\}$ respectively.

2.3 Essentials of set theory

(B&E, Appendix A)
Because events are subsets of the sample space, we can use the notation and theorems of set theory. Sets are usually indicated by capital letters, starting with *A.* The universal set contains all elements under consideration; so within the context of probability theory, this set is equal to the sample space *S.* To specify which elements are part of a certain set, we can explicitly list each element in the set, for example $A = \{2, 4, 6\}$, or $S = \{1, 2, 3, 4, 5, 6\}$. If *a* denotes an element of the set *A,* we write $a \in A$, so $2 \in \{2, 4, 6\}$. We can also describe the set, as in "*A* consists of all even numbers of dots when throwing a die"; or, more formal, $A = \{ x \mid x \in S \text{ and } x \text{ is even } \}$, where the bar "|" should be read as "such that".
If all elements in a set *A* are also elements of another set *B*, we say that *A* is a subset of *B*, notated as $A \subset B$, for example: $\{ 2, 4 \} \subset \{ 2, 4, 6 \}$. A set which does not contain any element is called the empty set, and is denoted by the symbol $\varnothing$.

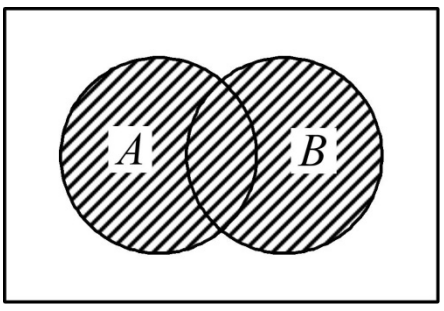


A Venn-diagram can be very helpful. The points inside the rectangle represent the sample space *S.* Events are represented by areas within the rectangle, usually circles (see *A* and *B*).



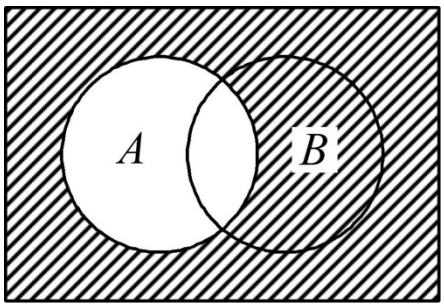
$A \cap B$, the **intersection** of *A* and *B.*
(Dutch: doorsnede).
Say: '*A* and *B*'.

Defined as: $A \cap B = \{x \mid x \in A \text{ and } x \in B\}$
E.g.: If $A = \{1, 2, 3\}$ and $B = \{2, 4, 6\}$, then $A \cap B = \{2\}$.

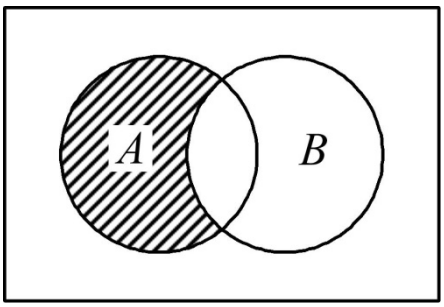


$A \cup B$, the **union** of *A* and *B.*
(Dutch: vereniging).
Say: '*A* or *B*'.

Defined as: $A \cup B = \{x \mid x \in A \text{ or } x \in B\}$
E.g.: If $A = \{1, 2, 3\}$ and $B = \{2, 4, 6\}$, then $A \cup B = \{1, 2, 3, 4, 6\}$.

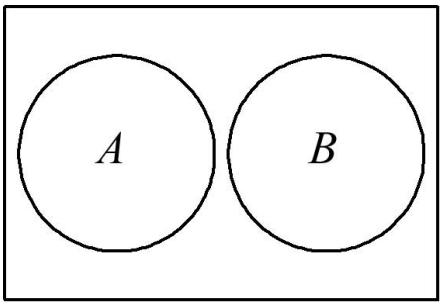


$\overline{A}$ or A' or A^c, the **complement** of A.
Say: 'Not A'.
Defined as: $\overline{A} = \{x \mid x \in S \text{ and } x \notin A\}$
E.g.: If $A = \{1, 2, 3\}$ and $S = \{1, 2, 3, 4, 5, 6\}$, then $\overline{A} = \{4, 5, 6\}$



$A \cap \overline{B} = A - B$, the **difference** of A and B.
(Dutch: verschil).
Say: 'A minus B'.

E.g.: If $A = \{1, 2, 3\}$ and $B = \{2, 4, 6\}$, then $A \cap \overline{B} = \{1, 3\}$



$A \cap B = \varnothing$. If A and B have no elements in common, they are **disjoint** or **mutually exclusive.**
(Dutch: disjunct of elkaar uitsluitend).
E.g.: $A = \{1, 2, 3\}$ and $B = \{4, 6\}$ are disjoint.

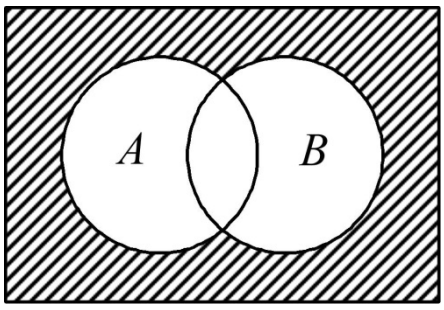
All of the above can be extended in a trivial way to situations with more than two subsets. We give without proof a few basic theorems, which can be checked easily with the use of Venn-diagrams:

1. Commutative laws: $A \cup B = B \cup A$ and $A \cap B = B \cap A$
2. Associative laws: $A \cup (B \cup C) = (A \cup B) \cup C$ and $A \cap (B \cap C) = (A \cap B) \cap C$
3. Distributive laws:
   $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$ and $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$

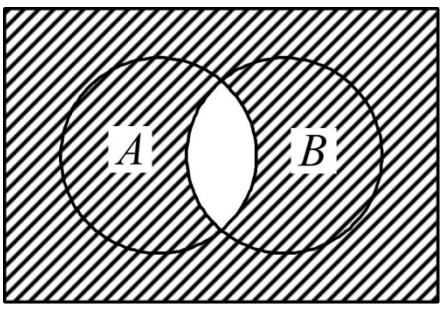
The following theorem is also sometimes useful in probability problems:

Theorem 2.1 (De Morgan-laws)

$\overline{(A \cup B)} = \overline{A} \cap \overline{B}$ and $\overline{(A \cap B)} = \overline{A} \cup \overline{B}$



$\overline{(A \cup B)} = \overline{A} \cap \overline{B}$



$\overline{(A \cap B)} = \overline{A} \cup \overline{B}$

2.4 Probability definitions and basic probability rules

(B&E, page 7-15)
There are three different ways of looking at the concept 'probability', each one intuitive and simple. These will be discussed first, followed by an axiomatic approach. This latter approach will then be used to derive a number of basic probability rules.

2.4.1 Classical probability (Laplace)

The first successful attempts to create a mathematical theory for the calculation of probabilities were particularly motivated by games of chance. Using contributions of many others, Laplace formulated in the 19-th century a definition of probability which came later to be known as the classical definition. When we throw a die, where each of the number of dots (1 to 6) is equally likely (i.e. an unbiased die), then it is very straightforward to say that the probability of throwing 3 dots is equal to 1/6. The same probability also applies to the other 5 possible outcomes. In this way we can determine probabilities simply by counting the number of outcomes leading to a particular event. For example, the event of an even number of dots contains the three different outcomes 2, 4, 6, and therefore the probability for this event is 3 divided by 6, or ½.
More generally: the *probability of an event* is the number of outcomes in that event divided by the total number of possible outcomes, provided that all the outcomes are 'equally likely'. We denote the total number of outcomes by $n(S)$ and the total number of those that are within event A by $n(A)$. Then the probability $P(A)$ of A can be written as:

$$P(A) = \frac{n(A)}{n(S)}$$

Note that it is possible to determine such a probability without actually performing a random experiment. The term 'equally likely' refers to the absence of any preference for any of the outcomes in the sample space; often, it is said that the sample space is **symmetrical** in that case.

Example 2.2

An *unbiased coin* (or fair; Dutch: zuivere munt) is defined as a coin where both sides (H=heads and T=tails) have an equal probability of showing up when tossing the coin. Applying the classical definition to this experiment, we can say for example that the probability of Heads is equal is ½. When tossing an unbiased coin *twice*, we could write the sample space as { HH, HT, TH, TT }. Again, each of these four outcomes is equally likely, so for example we can say that the probability of 'HH' is equal to ¼. But note: when we count the number of times that Heads show up, we can also define the sample space as S = { 0, 1, 2 }. Now the outcomes in S are no longer equally likely (S is no longer symmetrical), so this definition of the sample space cannot be used to find probabilities with the classical approach. This trivial example stresses the importance to consider the question whether it is reasonable to assume that all outcomes are equally likely. ◄

Using this same principle and employing proper counting techniques (see section 2.7), we can easily calculate probabilities in much more complicated situations, like for example the probability that someone playing bridge (a card game) will be dealt 6 spades, 2 clubs, 3 diamonds and 2 hearts.
But there are also many questions that cannot be answered with the classic approach. For example, what should we do if an infinite number of outcomes are possible? And, of course, this definition is completely impossible when the sample space is not symmetrical. For those cases, we need another definition of probability.

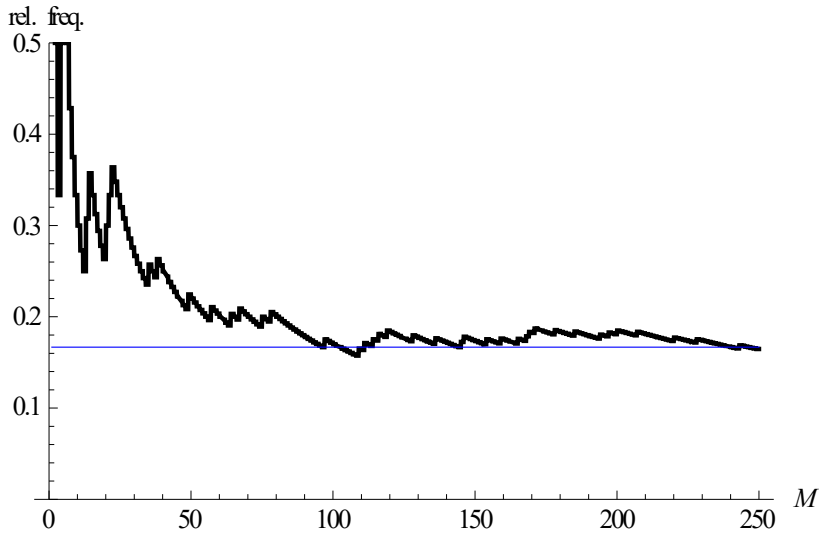
2.4.2 Probability as relative frequency

If we do not know whether a coin is unbiased or not, we cannot use the classical approach to determine the probability of 'Heads'. However, it is still a useful concept to talk about the probability of tossing 'Heads'. Now think of a situation where this coin will be tossed many times. When this number of tosses approaches infinity, we can expect the proportion of times a head occurs to converge to a certain constant p. So p will be equal to the relative frequency of heads showing up in the very long run. We can define the probability of 'Heads' as this limiting value p.

Formally: an experiment is performed M times. When we denote the number of times that an event A occurs by $m(A)$, then we *define* the probability of A as:

$$P(A) = \lim_{M \to \infty} \frac{m(A)}{M}$$

As an illustration we will use this definition to find the probability of 'three dots' when throwing a particular die. The graph below shows the relative frequency of 'three dots' we might observe as a function of the number of times the die has been rolled. When we continue to roll the die, we will see that the relative frequency will converge more and more to a certain value. We can see from the graph that it seems that this particular die may very well have been an unbiased one, simply because it looks like the relative frequency converges to 1/6 (indicated by horizontal line).



It is important to stress that we will never be able to find an exact probability using this approach. No matter how many times we will roll the die, the result will always be just an estimate of this true (but unknown) probability. During the next course, Probability Theory and Statistics 2, we will be able to draw some conclusions about the quality of such an estimate, based on the number of times the die has been rolled.
The convergence as described above is a consequence of the Law of Large Numbers, which will also be discussed at a later stage (Probability Theory and Statistics 3).

2.4.3 Subjective probability

We can also speak about probabilities in situations where both the classical approach as the relative frequency approach do not lead to results. For example, what is the probability that the third world war breaks out before the year 2050? Every individual can assign a different value to that probability. But the same general probability rules which will be discussed in the next sections will be applicable to these subjective probabilities as well.

2.4.4 Axioms of probability theory

For a long time probability theory was based on the classical approach, by defining sample spaces with a finite number of equally likely outcomes. Artificial 'tricks' were developed in order to model situations where the sample space was either infinitely large or not symmetrical in such a way that they fitted the classical framework. More and more, this lead to insurmountable problems in the theory. Only in 1934, the Russian mathematician Kolmogorov proposed a solid mathematical framework. Like any mathematical theory, he started by defining the axioms of probability theory.

To do so, we consider a set function $P(\cdot)$, that assigns a real value to each event A which is a subset of the sample space S. The function $P(\cdot)$ is called a **probability set function** (Dutch: **kansmaat**), and $P(A)$ is called the **probability** of A, if the following three properties (axioms) are satisfied:

1. $P(A) \geq 0$ for every event A (a probability is never negative).
2. $P(S) = 1$ (the probability of observing an outcome within S equals 1).
3. If $A_1, A_2, \ldots$, is a sequence of mutually exclusive events (i.e. $A_i \cap A_j = \varnothing$ for $i \neq j$, then:

   $$P(A_1 \cup A_2 \cup A_3 \cup \ldots) = \sum_{i=1}^{\infty} P(A_i).$$

   (In words: for any sequence of events which cannot occur at the same time, you can find the probability that any one of these events will occur by summing the probabilities for each of the individual events.)

The intuitive concepts of probability, both the classical as well as the relative frequency approach, satisfy these axioms. For example, we can see that axiom 1 follows directly from both probability definitions. The same applies to the condition that the total probability must be equal to 1. And the probability that someone throws an even number of dots is, again according to both definitions, clearly equal to the sum of the probability of a '2' and the probability of a '4' and the probability of a '6' (axiom 3).
Note that these axioms do not tell us what value the probability function $P(A)$ assigns to any given event A. Nevertheless, in combination with the definitions which will be introduced step-by-step, they do give a firm foundation for all theorems and proofs in probability theory. This starts at a very basic level:

Theorem 2.2

$P(\varnothing) = 0$

Proof

This theorem sounds trivial, but how can we prove it using the axioms?
Set $A_1 = \varnothing$, $A_2 = \varnothing$, $A_3 = \varnothing$, etc. Then it is clear that this sequence of events is mutually exclusive and that $A_1 \cup A_2 \cup A_3 \cup \ldots = \varnothing$. So we can apply axiom 3, resulting in:

$$P(\varnothing) = P(A_1 \cup A_2 \cup A_3 \cup \ldots) = \sum_{i=1}^{\infty} P(A_i) = \sum_{i=1}^{\infty} P(\varnothing)$$

It can easily be seen that this equality can only hold in case $P(\varnothing) = 0$.

Theorem 2.3

If A and B are mutually exclusive events, then:

$$P(A \cup B) = P(A) + P(B). \qquad \text{(special addition rule)}$$

Proof Do it yourself!

Theorem 2.4 *(B&E, Th. 1.4.1)*

If A is an event and $\overline{A}$ is its complement, then:

$$P(\overline{A}) = 1 - P(A) \qquad \text{(complement rule)}$$

Proof

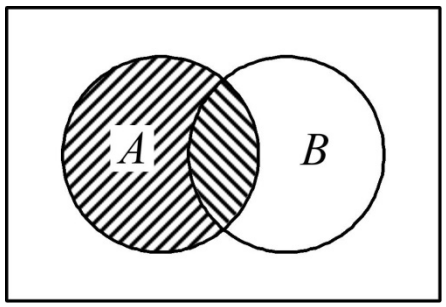
Because $\overline{A}$ is the complement of A, we know from set theory that $\overline{A} \cup A = S$ and $\overline{A} \cap A = \varnothing$. Therefore, we can apply Theorem 2.3 which gives $P(S) = P(\overline{A} \cup A) = P(A) + P(\overline{A})$. The result then follows from axiom 2.

Theorem 2.5

If A and B are any two events, then

$$P(A) = P(A \cap B) + P(A \cap \overline{B})$$

Proof



From set theory, we know that $A = (A \cap B) \cup (A \cap \overline{B})$ where $(A \cap B)$ and $(A \cap \overline{B})$ are disjoint (because $(A \cap B) \cap (A \cap \overline{B}) = \varnothing$).
So from Theorem 2.3 it follows that $P(A) = P(A \cap B) + P(A \cap \overline{B})$

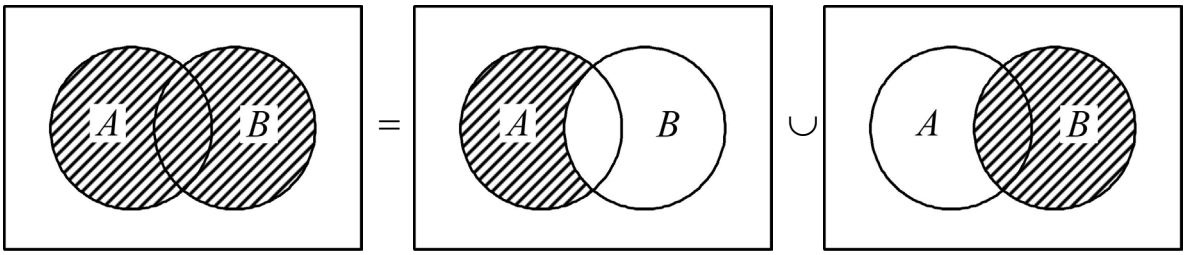
Theorem 2.6 *(B&E, Th. 1.4.3)*

If A and B are any two events, then

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$ (*general addition rule*)

Proof

As illustrated below by Venn diagrams, we know that $A \cup B = (A \cap \overline{B}) \cup B$:



It is simple to see that $(A \cap \overline{B})$ and B are mutually exclusive. Applying Theorem 2.3 we get

$$P(A \cup B) = P(A \cap \overline{B}) + P(B)\,.$$

From Theorem 2.5 it follows that the first term on the right-hand side of the equation above is equal to $P(A) - P(A \cap B)$, which completes the proof.

As usual, it helps a lot to look at Venn diagrams. $A \cup B$ contains all outcomes which belong to the union of A and B exactly once. In the addition $P(A) + P(B)$, the elements in the intersection of A and B are counted twice, which explains why we still need to subtract $P(A \cap B)$.

2.5 Conditional probability and independence

(B&E, page 16-28)

2.5.1 Conditional probability

Consider the experiment of randomly selecting two balls, one after the other, from a box with 5 black and 4 white balls. What is now the probability that both balls are black? Many people, even those not trained in probability theory, will probably more or less know how to tackle this problem. It seems reasonable to start by determining the probability that the first ball will be black (5/9). Next, one would determine the probability that the second ball will also be black (4/8, because the box will then

contain only 8 balls, of which 4 are black). Then one could multiply these two probabilities, such that the (correct!) answer would be: (5/9)*(4/8) = 5/18. Without properly realising it, the concept of *conditional probability* has been used here already, i.e. the probability that the second ball is black *given that the first ball was black*. We can write:

P("1st black" and "2nd black") = P("1st black") * P("2nd black given that 1st black").

Instead of writing each time the word 'given', we will use the vertical bar, so the latter probability will be written as P("2nd black | 1st black").
When we use the notation A and B for the events, we can write:

$$P(A \cap B) = P(B)P(A \mid B) \quad \text{or} \quad P(A \cap B) = P(A)P(B \mid A) .$$

In order to build up probability theory in a formal way, we will need definitions in addition to the axioms and theorems. Most books start the discussion about conditional probabilities by giving the following definition, which is essentially exactly the same as what we derived above.

Definition 2.1 *(B&E, Def. 1.5.1)*

The **conditional probability** (Dutch: voorwaardelijke of conditionele kans) of event A given the event B (say: probability of A given B) is defined by

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}, \quad \text{if } P(B) > 0$$

We will illustrate this formula with the following example.

Example 2.3

At a faculty of a certain university, students ({Male, Female}) can enrol in exactly one of the three offered programmes {E, F, G}. The student population is as follows.

	E	F	G	total
Male	1050	900	1200	3150
Female	900	450	1500	2850
Total	1950	1350	2700	6000

Before we can talk about probabilities, we have to define an experiment: we select at random one student from the population above. The sample space consists of $n(S)$=6000 elements, each equally likely to be selected. Using the classical probability definition, we can find the following table with probabilities (all numbers from the first table are divided by 6000):

	E	F	G	total
Male	0.175	0.150	0.200	0.525
Female	0.150	0.075	0.250	0.475
Total	0.325	0.225	0.450	1.000

So the probability that a randomly selected student is a male who is enrolled in E is 0.175. Now what is the conditional probability that a student is male given that the student is enrolled in E? In determining that probability we can limit ourselves to only those students who are enrolled in E. Of those 1950 students, 1050 are males and 900 are females. So it makes sense to state that the requested conditional probability is equal to

$$P(\text{Male} \mid \text{E}) = \frac{1050}{1950} = \frac{n(\text{Male} \cap \text{E})}{n(\text{E})} = 0.538$$

But this probability is also equal to:

$$P(\text{Male} \mid \text{E}) = \frac{n(\text{Male} \cap \text{E}) / n(S)}{n(\text{E}) / n(S)} = \frac{P(\text{Male} \cap \text{E})}{P(\text{E})} = \frac{0.175}{0.325} = 0.538$$

Note that this corresponds exactly to Definition 2.1. Note as well that the two probabilities $P(\text{Male} \mid \text{E})$ and $P(\text{E} \mid \text{Male})$ are very different! ◄

When we are dealing with conditional probabilities given an event *B*, it means that we assume the random experiment has led to an outcome in *B*. Therefore, *B* can be seen as our 'new' sample space. It can be easily shown that the previously derived results for probabilities in relation to the sample space *S*, still remain valid when the sample space is now limited to *B*. For example:

$$P(S \mid B) = \frac{P(S \cap B)}{P(B)} = \frac{P(B)}{P(B)} = 1$$

$$P(\overline{A} \mid B) = 1 - P(A \mid B)$$

$$P(A_1 \cup A_2 \mid B) = P(A_1 \mid B) + P(A_2 \mid B) - P(A_1 \cap A_2 \mid B)$$

(see exercises!)

The definition of conditional probability leads directly to the following result, which we derived already at the beginning of this section.

Theorem 2.7 (Multiplication theorem) *(B&E, Th. 1.5.1)*

If *A* and *B* are any two events with $P(A) > 0$ and $P(B) > 0$), then

$$P(A \cap B) = P(A) \cdot P(B \mid A) = P(B) \cdot P(A \mid B)$$

By applying this theorem twice, we arrive at the following result (check for yourself!):

$$P(A \cap B \cap C) = P(A) \cdot P(B \mid A) \cdot P(C \mid A \cap B)$$

2.5.2 Independent events

If knowing that event *A* has occurred does not affect the probability of *B* occurring, then we can write $P(B \mid A) = P(B)$. In this case, it is straightforward to say that event *B* is independent of event *A*. But using Theorem 2.7, we can see that this is equivalent to writing: $P(A \cap B) = P(A) \cdot P(B)$. And that is exactly how most text books define independent events:

Definition 2.2 *(B&E, Def. 1.5.2)*

Two events *A* and *B* are called *independent* (Dutch: *onafhankelijk*) if

$$P(A \cap B) = P(A) \cdot P(B)$$

Compare this equation with the equation from Theorem 2.7. Combining Definition 2.1 and Definition 2.2 we get (if $P(B) > 0$ and *A* and *B* independent):

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A) \times P(B)}{P(B)} = P(A)$$ (only if *A* and *B* are independent)

and vice versa: $P(B \mid A) = P(B)$ (if $P(A) > 0$).

Example 2.4

Consider the experiment of drawing one card from a deck of 52 playing cards. Define *A* as drawing a spade, and *B* as drawing a knight. Using the classical probability approach, it is easy to see that $P(A) = 1/4$ and $P(B) = 1/13$. Also $P(A \cap B) = 1/52$, since only one card is both a spade and a knight. This confirms the equality between $P(A \cap B)$ and $P(A) \cdot P(B)$ in this example, or, in other words: *A* and *B* are independent events. Alternatively, we could have drawn the same conclusion by first determining $P(B \mid A)$. This probability is 1/13, since of all 13 spades, exactly one is a knight as well. It follows that $P(B \mid A) = P(B)$, so *A* and *B* are again shown to be independent. Finally, we can

also see that $P(A \mid B) = P(A)$ $(= 1/4)$. These three ways to check for independence are always equivalent. ◄

Example 2.5

Look again at Example 2.3. Are the events 'Male' and 'E' independent? The answer is negative, because we have seen before that *P*(Male) = 0.525, while *P*(Male | E) = 0.538. In other words, knowledge of the fact that a student is enrolled in E does change the probability that the student is Male. Similarly, we can also see that *P*(E | Male) = 0.333, which is not equal to *P*(E) = 0.325. ◄

Beginners sometimes confuse the concepts 'disjoint' and 'independent' events. In general we can say that disjoint events are NOT independent! Because if *A* and *B* are disjoint events, they have an empty intersection; if event *B* occurs, we can be sure that the event *A* did not occur, so $P(A \mid B) = 0$. But for *A* and *B* to be independent, $P(A \mid B)$ must be equal to $P(A)$, which for disjoint events is possible only when $P(A) = 0$. However, we are usually not interested in events that can never occur, so in general the concepts 'disjoint' and 'independent' have a very different meaning.

2.5.3 Law of total probability and Bayes' theorem

In set theory, a **partition of a set** is a classification of the set's elements into non-empty subsets, in such a way that every element is included in one and only one of the subsets. Applied to probability theory: $\{E_1, E_2, \ldots, E_k\}$ is a partition of the sample space *S* if $E_1, E_2, \ldots, E_k$ are mutually exclusive events whose union is *S* (B & E uses the phrase "mutually exclusive and exhaustive events"). We will consider here only partitions where the number of subsets is finite; the subsets $E_1, E_2, \ldots, E_k$ are sometimes also called classes or categories, and we will consider here only partitions with a finite number of categories.

Example 2.6

The collection of elementary outcomes forms a partition, so when a die is thrown, the set {{1}, {2}, {3}, {4}, {5}, {6}} forms a partition of the sample space into 6 subsets. But many other partitions are possible, for example { 'even', 'odd'} = { {2,4,6}, {1,3,5} }. Or, if we throw three dice, we could calculate the sum of the dots, and define the next partitions into three categories each: { 3 to 6, 7 to 12, 13 to 18}, or {3 to 12, 13 to 16, 17 to 18} ◄

A **cross table** (Dutch: kruistabel) splits the sample space into two different partitions. If the first partition counts *k* categories, and the other *r*, then there are $k \times r$ different pairs of categories. Say we have a certain event *A*, then $\{ A, \overline{A} \}$ forms a partition with two categories. Assume also another partition $\{E_1, E_2, \ldots, E_k\}$. A cross table with the so-called simultaneous probabilities ($P(A \cap E_1)$ etc.) can be displayed as:

	E_1	E_2	….	E_k
A	$P(A \cap E_1)$	$P(A \cap E_2)$	….	$P(A \cap E_k)$
not A ($\overline{A}$)	$P(\overline{A} \cap E_1)$	$P(\overline{A} \cap E_2)$	….	$P(\overline{A} \cap E_k)$

From a table like this, we can easily derive probabilities like $P(A)$, $P(\overline{A})$, $P(E_1)$. For example, it is clear that:

$$A = (A \cap E_1) \cup (A \cap E_2) \cup \cdots \cup (A \cap E_k)$$

Because all events $A \cap E_1$, $A \cap E_2, \ldots, A \cap E_k$ are disjoint (which follows directly from the fact that $E_1, E_2, \ldots, E_k$ are disjoint), we can apply Theorem 2.3:

$$P(A) = \sum_{i=1}^{k} P(A \cap E_i)$$

These probabilities can be calculated in the 'margins' of the cross table, simply by summing all simultaneous probabilities in each row and each column. The resulting probabilities are then usually referred to by the term **marginal probabilities**.

	E_1	E_2	E_k	Total
A	$P(A \cap E_1)$	$P(A \cap E_2)$	$P(A \cap E_k)$	$P(A)$
not A ($\overline{A}$)	$P(\overline{A} \cap E_1)$	$P(\overline{A} \cap E_2)$	$P(\overline{A} \cap E_k)$	$P(\overline{A})$
Total	$P(E_1)$	$P(E_2)$	$P(E_k)$	1

In Example 2.3 we already encountered such a cross table. There, the student population was split according to two different partitions, by programme (three categories E, F and G) and by gender (two categories Male, Female).

Theorem 2.8 Law of Total Probability *(B&E, Th. 1.5.2)*

If A is any event and $\{E_1, E_2, \ldots, E_k\}$ is a partition of S, then:

$$P(A) = P(A \mid E_1) \cdot P(E_1) + P(A \mid E_2) \cdot P(E_2) + \ldots + P(A \mid E_k) \cdot P(E_k) = \sum_{i=1}^{k} P(A \mid E_i) \cdot P(E_i)$$

Proof

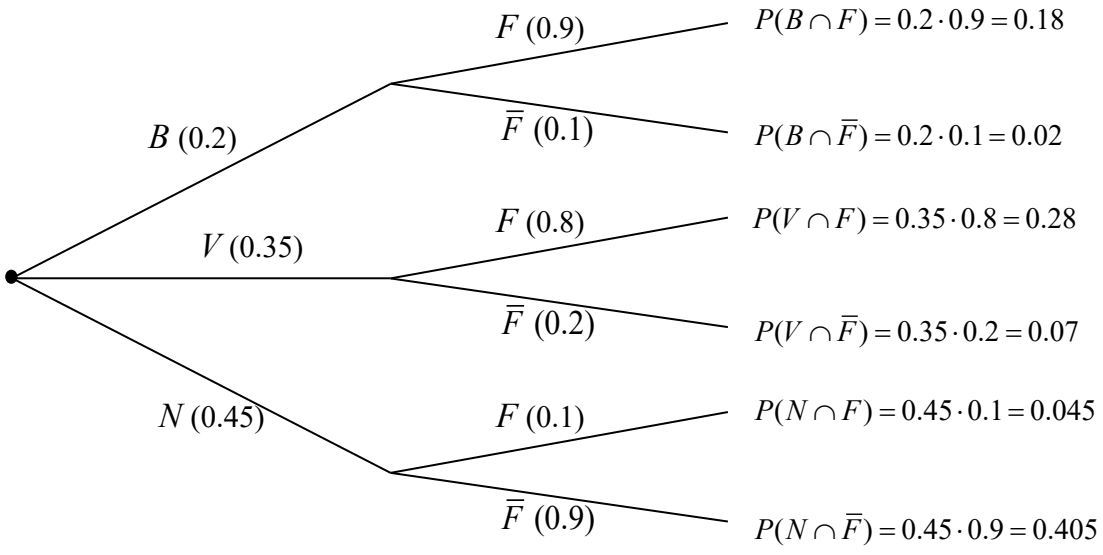
All events $A \cap E_1$, $A \cap E_2, \ldots, A \cap E_k$ are mutually exclusive, so $P(A) = \sum_{i=1}^{k} P(A \cap E_i)$.

We know from Theorem 2.7 that $P(A \cap E_i) = P(A \mid E_i) \cdot P(E_i)$ for each i, which gives the result.

Tree diagrams (Dutch: kansbomen) can be a helpful tool in these situations. From the origin we can draw branches which correspond to the classes for one partition, and from the endpoint of each of these branches we draw the branches according to the other partition (and from those endpoints, even branches for a possible third partition can be drawn). Along the branches, we write the (conditional) probabilities.

Example 2.7

A patient visits his general practitioner. Assume that any arbitrary patient has either a bacterial infection (B), ór a viral infection (V), óf no infection (N) with P(B) = 0.20, P(V) = 0.35 and P(N) = 0.45 (note that B, V and N are thus mutually disjoint). Also assume that the probability a patient has a fever (F) depends on the type of infection he has: P(F | B) = 0.90, P(F | V) = 0.80 and P(F | N) = 0.10. What is the probability that an arbitrary patient has a fever? A tree diagram looks like:



According to the Law of Total Probability, we get

$$P(F) = P(B)P(F \mid B) + P(V)P(F \mid V) + P(N)P(F \mid N) = 0.18 + 0.28 + 0.045 = 0.505 .$$

This is shown in the tree diagram by summing the probabilities at the right-hand side for all the branches that pass an 'F'. ◄

It often happens that certain conditional probabilities are given (like $P(A \mid E_i)$), but that actually the reverse conditional probabilities are requested (such as $P(E_i \mid A)$). The following theorem is very important:

Theorem 2.9 (Bayes' Theorem) *(B&E, Th. 1.5.3)*

If A is any event and $\{E_1, E_2, \ldots, E_k\}$ is a partition of S, then:

$$P(E_i \mid A) = \frac{P(A \mid E_i) \cdot P(E_i)}{P(A \mid E_1) \cdot P(E_1) + P(A \mid E_2) \cdot P(E_2) + \ldots + P(A \mid E_k) \cdot P(E_k)}$$

If we choose $\{B, \overline{B}\}$ as a partition, then this simplifies to:

$$P(B \mid A) = \frac{P(A \mid B) \cdot P(B)}{P(A \mid B) \cdot P(B) + P(A \mid \overline{B}) \cdot P(\overline{B})}$$

Proof

From Definition 2.1 and Theorem 2.7, we know that

$$P(E_i \mid A) = \frac{P(A \cap E_i)}{P(A)} = \frac{P(A \mid E_i) \cdot P(E_i)}{P(A)}$$

We obtain the desired result after replacing the denominator in the last expression by the result of Theorem 2.8.

Example 2.8

Using the previous example, we will now answer the following question: what is the probability that someone who has a fever, is actually suffering from a bacterial infection? We use Bayes' rule applied to $P(B \mid F)$ to find:

$$P(B \mid F) = \frac{P(F \mid B) \cdot P(B)}{P(F \mid B) \cdot P(B) + P(F \mid V) \cdot P(V) + P(F \mid N) \cdot P(N)} = \frac{0.18}{0.505} = 0.3564$$

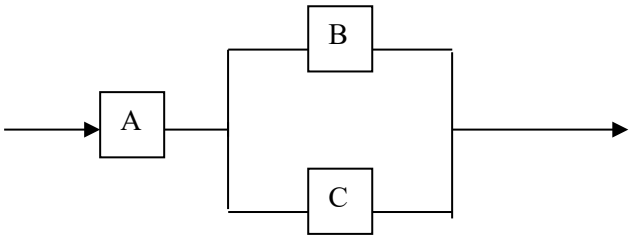
Note that the probability of a patient having a bacterial infection has increased from 0.2 to 0.356 as a result of the fact that some additional information has become available, namely that the patient has a fever. This is in fact similar to the way any General Practitioner (GP) works, since the GP will, usually subconsciously, update his assessment of the most likely disease after a new piece of information has become available. ◄

2.6 Application: reliability of systems

As an application of probability rules, we will discuss simple reliability calculations of systems that are made up of components, each of which may fail with a certain (known) probability. We will make a distinction between components connected in series or connected in parallel. In a parallel configuration, the system can still operate even if one component has become defective. In a series configuration, the system will fail as soon of one of the components fails. For all components, it will be assumed that failure is independent of whether or not other components have failed. An example will clarify most of these ideas.

Example 2.9

A system with three components may be represented by the following chart:



We say that the system still functions as long as a flow is possible through the system from left to right. So in the flow chart above, in order for the system to function, it is essential that A functions and either B or C (or B and C both). The components B and C are *parallel-connected* components, and the group (B, C) is *connected in series* with component A. If A is faulty, then the system fails as a whole, just as is the case when both B and C are defective. ◄

Each component is characterised by a certain probability that it will continue to work (during some fixed time interval), which is called the *reliability* of that part. Similarly, each component has a certain fail probability, which is of course simply 1 minus the reliability. With this information for each component, we can determine the reliability of the system as a whole, which is the probability that the system does not fail (during the time interval). We continue with the example.

Example 2.10 *(continuation of previous example)*

We assume that the probability that A fails is 0.017, and that those probabilities for B and C are equal to 0.2 and 0.15. The probability that the group B-C works, is calculated with the complement rule and special multiplication rule. The probability that the group *fails*, is equal to the probability that both B and C fail, which is $0.2 \times 0.15 = 0.03$. The probability that the group B-C continues to function, is $1 - 0.03 = 0.97$. For the operation of the system as a whole, group B-C should not fail and A should not fail. With the special multiplication rule (independence of events!), the reliability of the system is found: $0.97 \times (1 - 0.017) = 0.97 \times 0.983 = 0.95351$. The probability that the system fails, is $1 - 0.95351 = 0.04649$.

We can perform similar calculations even for much more complicated systems, for example, by replacing the blocks A, B and C by compound systems. It is also possible to require, for example, that of three parallel components at least two should remain functional. ◄

2.7 Counting techniques (combinatorics)

(B&E, page 31-42)

As we have seen before, in experiments where all outcomes are equally likely, we can express a probability as a ratio, where both numerator and denominator represent 'the number of outcomes' in the event and in the sample space respectively. It is therefore very useful if we are able to quickly determine such numbers. Some counting techniques can help us.

2.7.1 The multiplication principle

The **multiplication principle** is, stated simply, the idea that if there are n_A ways of doing something and n_B ways of doing another thing, then there are $n_A \times n_B$ ways of performing both actions.

Example 2.11

If Pete can choose between five shirts and three pairs of trousers, then he can dress himself in $5 \times 3 = 15$ different ways. ◄

Example 2.12

This principle can easily be extended to more than two actions. For example, the total number of five-digit numbers where each digit is selected from the set {1, 2, ..., 7}, is equal to $7 \times 7 \times 7 \times 7 \times 7 = 7^5$ (this is an example of drawing with replacement). ◄

Even in cases where the different choices themselves may vary, *but not their number*, this principle can be applied.

Example 2.13

A box contains 10 balls numbered 1, 2,. . . , 9, 10. We will randomly select balls one after the other without putting the previously drawn balls back into the box (drawing without replacement). Of course, it depends on the first ball drawn which balls will still be available for the second draw. However, it is totally clear that the number of remaining balls equals nine. So if we select at random three balls without replacement from this box, there are $10 \times 9 \times 8 = 720$ ways of doing this. ◄

Apart from the difference between drawing with or without replacement, two other aspects are important when counting the number of ways. In the examples above, we assumed that the **order** in which the elements were drawn is relevant. In Example 2.13, we counted the number of ways assuming that the outcome (4, 2, 7) (in this particular order) is a different outcome from (7, 4, 2). The other aspect concerns the question whether all elements which can be selected can be **distinguished** from each other or not. If the box in Example 2.13 does not contain 10 balls with all different numbers, but for example four blue and six green balls, then we can say that the six green balls are mutually indistinguishable. Of course, this will reduce the number of different ways in which we can select three balls. These situations will now be discussed in more detail, where we will start the discussion with situations in which all elements are distinct, and order matters.

2.7.2 Permutations

A **permutation** is a particular arrangement of a collection of objects or numbers. Simple example: assume there are 4 marbles, red, yellow, green and blue. Thus, "red, yellow, green, blue" is a specific permutation, just as "red, green, yellow, blue". We now determine the number of possible arrangements for n distinct objects. If n is not too large, one could simply list all of these sequences, for example, for $n = 3$:

a, b, c a, c, b b, c, a b, a, c c, a, b c, b, a

We can also see that three objects are available for the first position, two remaining objects for the second position and just one for the third position, which indeed gives $3 \times 2 \times 1 = 6$ ways. More generally for arbitrary values of n: n possibilities for the first position, $n - 1$ for the second, $n - 2$ for the third, and so on. Thus, $n! = n \times (n - 1) \times \ldots 2 \times 1$ is the number of permutations of n distinguishable objects. ($n!$ is ***n*-factorial** (Dutch: n-faculteit), a very rapidly increasing function of n, where $0! = 1$ by definition). For the four marbles above, 24 permutations exist.

We also use the term permutations when we do not consider arrangements of all n elements, but of only r elements ($r \le n$). (Note: Dutch books usually use the term "Variaties" whenever $r < n$). In Example 2.13 we have seen such a situation already. The number of permutations of r objects drawn from a set of n distinct objects can be derived in a similar manner, and is equal to $n \cdot (n-1) \cdot \cdots \cdot (n-r+1)$. This result can be written as well as:

$$ {}_nP_r = n \cdot (n-1) \cdot \cdots \cdot (n-r+1) = \frac{n!}{(n-r)!} \quad \text{(Check!)} $$

Example 2.14

The number of permutations when drawing without replacement three balls from a box with balls numbered from 1 t / m 10, is ${}_{10}P_3 = 10 \cdot 9 \cdot 8 = \frac{10 \cdot 9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = \frac{10!}{7!} = 720$ ◄

2.7.3 Combinations

If the order is irrelevant, then we can talk about the number of different subsets which can be the result when drawing k objects from a set of n distinguishable objects.

Example 2.15

Again, we draw three numbers without replacement from the set { 1, 2,. . . , 10 }, but now we are no longer interested in the order in which those numbers were drawn. In Example 2.14, the number of permutations of three numbers from a collection of 10 numbers was found to be 720. But when the order is no longer relevant, we would regard the following six outcomes as essentially the same: (2,4,7), (2,7,4), (4,2,7), (4,7,2), (7,2,4) and (7,4,2), since each leads to the subset containing the numbers 2, 4 and 7. These six outcomes are of course the six different permutations of three different numbers. Note that the same holds true for *any* subset of three numbers. Thus, the number of $10 \times 9 \times 8 = 720$ permutations should be divided by $3! = 6$ to find the number of different subsets, so there are $\frac{10 \times 9 \times 8}{3!} = \frac{720}{6} = 120$ different subsets of three numbers selected from ten different numbers. Or, written alternatively:

$$\frac{10 \times 9 \times 8}{3!} = \frac{10 \times 9 \times 8 \times 7 \times \ldots \times 2 \times 1}{3! \times 7 \times \ldots \times 2 \times 1} = \frac{10!}{3! \times 7!}$$ ◄

For any nonnegative integer n and any nonnegative integer $k \leq n$, the number of **combinations** (subsets without sequence; Dutch: combinaties of grepen) when selecting k objects from a set of n distinct objects, is equal to

$$\frac{n \times (n-1) \times (n-2) \times \ldots \times (n-k+1)}{k!} =$$

$$= \frac{n \times (n-1) \times (n-2) \times \ldots \times (n-k+1) \times (n-k)!}{k! \times (n-k)!} = \frac{n!}{k! \times (n-k)!} = \binom{n}{k}$$

The number $\binom{n}{k}$ (also often notated as C_k^n) is the *binomial coefficient*, and can be read as "n choose k" (Dutch: "n boven k").

Whenever $k < 0$ or $k > n$, we define that $\binom{n}{k} = 0$.

Pocket calculators often have a button like $\boxed{n\ \mathbf{C}\ r}$ to calculate these binomial coefficients, such that

$\boxed{1}$ $\boxed{0}$ $\boxed{n\ \mathbf{C}\ r}$ $\boxed{3}$ $\boxed{=}$

will result in 120.

Whether or not we are dealing with a situation where the order is relevant, needs sometimes a bit of thought:

Example 2.16

Out of a group of 40 students, a committee consisting of three students should be formed. The committee does not have any other formal structure. The number of different committees is therefore equal to the number of combinations, so $\binom{40}{3} = \frac{40 \cdot 39 \cdot 38}{3 \cdot 2 \cdot 1} = 9880$.

However, if from the same group of students a student board has to be formed, with a chairman, secretary and treasurer, then the number of different student boards is $40 \cdot 39 \cdot 38 = 59280$. ◄

The following example shows how we can use the calculation of the number of combinations to determine probabilities.

Example 2.17

Again we will look at Example 2.15. Now, we will order the three numbers drawn from small to large. What is the probability that the number in the middle is '4'? (Note: we use now the *order of magnitude*, not the order of drawing). We answer this question by using the classical probability definition:

$$P(\text{'4' in the middle}) = \frac{\text{number of combinations with '4' in the middle}}{\text{number of combinations}}$$

We already know that the number of combinations in the denominator is 120. The numerator is also easy to determine: to the left of the 4 in the middle, there is a choice of 1, 2, or 3 (3 possibilities); the middle digit must be 4, so that can be done in only one way, and the largest number must be greater than 4, which gives 6 possibilities (5, 6, 7, 8, 9 or 10). Thus, the numerator is $3 \cdot 1 \cdot 6 = 18$, and the requested probability is 18/120, so 0.15.
Let us compare this with the probability that the *second digit drawn* is '4'. The denominator will be 720 (the number of permutations). The numerator can be found as follows: for the first number we have a choice out of nine numbers (all except '4'), the middle number must be '4', and for the last digit we have a choice out of the remaining eight numbers, so the numerator is $9 \cdot 8 = 72$. This results in the probability 72/720 = 0.1. (Check that this result could also have been found in a much more straightforward way!) ◄

Theorem 2.10

$$\binom{n}{k} = \binom{n}{n-k}$$

Proof Very simple by replacing the binomial coefficient, try yourself.

This theorem is also very logical: a selection of k objects from n objects, defines exactly the remaining $n-k$ objects. So selecting k objects is essentially the same as selecting the other $n-k$ objects! Realising this can make the calculation by hand of binomial coefficients quite simple. For example, $\binom{40}{37} = \binom{40}{3}$ and we find this coefficient by multiplying *three* decreasing factors in the numerator, starting with 40, and writing in the denominator 3!:

$$\binom{40}{3} = \frac{40 \cdot 39 \cdot 38}{3 \cdot 2 \cdot 1}.$$

Theorem 2.11

$$\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}$$

Proof
Check by evaluation of the expressions.

Example 2.18
We will illustrate the correctness of the theorem above by using Example 2.16, where the number of committees is $\binom{40}{3}$. If one of the 40 students is Ben, then this total number of committees should be equal to the number of committees of which Ben is a member plus the number of committees of which Ben is a *not* a member. That first number is clearly equal to $\binom{39}{2}$, because the remaining two members of the committee should be selected from the other 39 students, while the number of committees without Ben is $\binom{39}{3}$. That proves that $\binom{40}{3} = \binom{39}{2} + \binom{39}{3}$. ◄

The last two theorems can also be recognised in **Pascal's triangle**, where all numbers represent binomial coefficients. The rows of Pascal's triangle are conventionally enumerated starting with row $n = 0$ at the top (the 0th row). The entries in each row are numbered from the left beginning with $k = 0$. Thus, in the fourth row, we find the numbers 1, 4, 6, 4 and 1, which are the binomial coefficients for $n = 4$ and k ranging from 0 to 4. The numbers in a row can be found by summing the two numbers diagonally above:

$$\begin{array}{ccccccccccc}
 & & & & & 1 & & & & & \\
 & & & & 1 & & 1 & & & & \\
 & & & 1 & & 2 & & 1 & & & \\
 & & 1 & & 3 & & 3 & & 1 & & \\
 & 1 & & 4 & & 6 & & 4 & & 1 & \\
1 & & 5 & & 10 & & 10 & & 5 & & 1 \\
1 & & 6 & & 15 & & 20 & & 15 & & 6 & & 1 \\
1 & & 7 & & 21 & & 35 & & 35 & & 21 & & 7 & & 1
\end{array}$$

In the sixth row we can find for example that 15 = 5 + 10, which is just another representation of Theorem 2.11 with $n = 6$ and $k = 3$: $\binom{6}{2} = \binom{6-1}{2-1} + \binom{6-1}{2} = \binom{5}{1} + \binom{5}{2}$.

Theorem 2.12 (Binomial Theorem) *(B&E, Ex. 1.6.8)*

Binomial coefficients also appear in the **Binomial Theorem** (Dutch: binomium van Newton):

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}$$

Proof

A formal proof for this theorem needs knowledge of the principle of mathematical induction. We will not discuss that here; instead we will just try to make this formula more understandable. Let us look at the expression above for increasing values of n and compare the coefficients for each of the terms with Pascal's triangle:

$$(a+b)^1 = \sum_{k=0}^{1} \binom{1}{k} a^k b^{1-k} = \binom{1}{0} a^0 b^1 + \binom{1}{1} a^1 b^0 = 1a + 1b$$

$$(a+b)^2 = \sum_{k=0}^{2} \binom{2}{k} a^k b^{2-k} = \binom{2}{0} b^2 + \binom{2}{1} ab + \binom{2}{2} a^2 = 1b^2 + 2ab + 1a^2$$

$$(a+b)^3 = \sum_{k=0}^{3} \binom{3}{k} a^k b^{3-k} = \binom{3}{0} b^3 + \binom{3}{1} ab^2 + \binom{3}{2} a^2 b + \binom{3}{3} a^3 = 1b^3 + 3ab^2 + 3a^2 b + 1a^3$$

$$(a+b)^4 = \sum_{k=0}^{4} \binom{4}{k} a^k b^{4-k} = \binom{4}{0} b^4 + \binom{4}{1} ab^3 + \binom{4}{2} a^2 b^2 + \binom{4}{3} a^3 b + \binom{4}{4} a^4 =$$
$$= 1b^4 + 4ab^3 + 6a^2 b^2 + 4a^3 b^1 + 1a^4$$

For example, the coefficients for $(a+b)^3$ are 1, 3, 3, and 1 respectively, as can be found in the third row of Pascal's triangle. We can also note that $(a+b)^3 = (a+b)(a+b)^2$, and since $(a+b)^2 = 1b^2 + 2ab + 1a^2$ we obtain:

$(a+b)^3 = (a+b)(1a^2 + 2ab + 1b^2) = (1a^3 + 2a^2 b + 1ab^2) + (1a^2 b + 2ab^2 + 1b^3)$. By summing the coefficients for terms with the same powers of a and b we arrive at the coefficients for (for example: the coefficient for $a^2 b$ is 2 + 1 = 3, or $\binom{2}{1} + \binom{2}{0} = \binom{3}{1}$, exactly corresponding to Theorem 2.11).

Or, more generally, using the above discussion about combinations: the left-hand side of the binomial consists of n factors, each equal to $a + b$. If we expand the left-hand side completely,

we get an expression consisting of 2^n terms, for example $(a+b)^3 = a^3 + a^2b + a^2b + a^2b + ab^2 + ab^2 + ab^2 + a^2$. But among those terms there will usually be a some with the same powers of *a* and *b*. If we count how many terms are equal to $a^k b^{n-k}$, we should realise that we will only arrive at the term $a^k b^{n-k}$ when, of the *n* factors $a + b$, we select *k* times *a* and $n - k$ times *b*. So now we can recognise that this is exactly equal to the number of combinations when selecting *k* objects (in this case *k* factors) from a total of *n* objects (factors).

Theorem 2.13 *(B&E, Ex. 1.6.9)*

$$2^n = \sum_{k=0}^{n} \binom{n}{k} = \binom{n}{0} + \binom{n}{1} + \binom{n}{2} + \ldots + \binom{n}{n}$$

Proof

The equality follows directly by applying the binomial theorem with $a = b = 1$.
Another way of looking at this equation is as follows: consider a collection of *n* distinct elements, and we want to determine the total number of different subsets that can be defined for this collection (where it does not matter how many elements are contained within a subset). The left-hand side of the above equality can be found by the following reasoning: for each of the *n* elements that are two possibilities: either it is within or it is not within a particular subset. Utilising the multiplication principle, we find that the number of possible subsets of a set of *n* elements is therefore equal to 2^n (The empty set and the collection itself are included in this number). The total number of possible subsets can also be found in a different way, i.e. by taking the sum of the number of subsets with zero elements (=1, the empty set only), the number of subsets with one element, the number of subsets with two elements, etc. Since we know that there are exactly $\binom{n}{k}$ different subsets with *k* elements, we obtain the expression on the right-hand side in the theorem.

2.7.4 Non distinguishable objects

Suppose we want to count in how many different ways we can arrange 10 balls. We have seen that this number equals 10! (= 3628800) if all the balls are distinguishable. But what if the box contains four blue and six red balls? We can solve this with the following idea. First, we give labels to the four blue and to the six red balls so that all 10 balls are distinguishable, for example in this way: B_1, B_2, B_3, B_4 and $R_1, R_2, R_3, R_4, R_5, R_6$. Among the 10! permutations, many are actually identical when we no longer make a distinction between each of the blue balls and each of the red balls. Say that positions 2, 5, 6 and 8 are occupied by blue balls (and so positions 1, 3, 4, 7, 9 and 10) are occupied by red balls. Now, those four blue balls B_1 t/m B_4 can be permuted in 4! different ways without changing the position of the blue balls within the sequence of all 10 balls (we are just interchanging blue balls). Similarly, the six red balls can be permuted in 6! ways. So after dividing 10! by both 4! and by 6!, we obtain the number of permutations for this case: $\frac{10!}{4!6!} = 210 = \binom{10}{4}$.

It is not a coincidence that we find the same binomial coefficient, which also applies to the number of ways of selecting 4 objects out of 10 distinct objects. In order to clarify this, we will look at the problem of finding the number of permutations of four blue and six red balls from another point of view. After arranging all 10 balls in a row, we can give each ball a sequence number from 1 to 10. For each permutation we have to count, the blue balls will be assigned a different set of sequence numbers. So in how many ways can we draw four sequence numbers to assign to the blue balls? The answer is clearly the number of combinations of drawing four objects (sequence numbers) out of a set of 10 sequence numbers, so indeed we obtain $\binom{10}{4}$ again.

More general: The number of permutation of n objects, of which r are of a certain type and $n - r$ are of a different type, is equal to $\binom{n}{r} = \frac{n!}{r!(n-r)!}$

With exactly the same argument, we can obtain an even more general result: The number of permutations of n objects, of which r_1 are of type 1, r_2 are of type 2,…. and r_k are of type k (with $n = r_1 + r_2 + ... + r_k$) is: $\frac{n!}{r_1!r_2!\cdots r_k!}$.

Example 2.19

Consider the 11 letters of the word 'abracadabra'. What is the probability of forming this word again when the letters are shuffled and put in an arbitrary order? There are 5 a's, 2 b's, 2 r's, 1 c and 1 d in this word, so the number of permutations is $\frac{11!}{5!\cdot 2!\cdot 2!\cdot 1!\cdot 1!} = 83160$. Of all these permutations only one leads to the formation of the word 'abracadabra', so the requested probability is 1/83160. ◄

2.8 Problems

2.1 Two balls are drawn from a box containing 9 balls that are numbered 1, 2, . . . , 9. Describe an appropriate sample space if (a) the first ball is not replaced before the second ball is drawn, (b) the first ball is replaced before the second ball is drawn.

2.2 In an electric circuit two switches 1 and 2 are connected in series. Switch 3 is parallel-connected with the group of switches 1 and 2. Each of the switches can be *open* or *closed*. If the switch is *closed* there can be an electric current, which is not possible if the switch is *open*. Describe an appropriate sample space that gives all possible positions of the three switches. In which positions there can be an electric current in the circuit?

2.3 A and N play a game called 'best-of-seven', in which a player wins if he has won at least four sets. List all possible outcomes in which A wins in a game with no more than 6 games.

2.4 Two numbers are drawn without replacement from the set {2, 3, . . . , 7}. The second number should be less than the first one, which means a new number will be drawn if the second one is higher. This process will continue until the second number is less, or until no numbers are left anymore. Describe the sample space.

2.5 A balanced coin is tossed five times and after each toss the cumulative numbers of 'heads' and 'tails' are recorded. Which outcomes do have the property that the number of 'heads' already tossed is at any stage (so after each of the five tosses) higher than the number of 'tails' already tossed?

2.6 Consider the examples B&E 1.2.1-1.2.4.
- a In which of these examples is the sample space symmetrical and in which is the sample space not symmetrical?
- b In which way you can extend the sample space in example 1.2.3 such that a symmetrical sample space arises? (of course infinitely large, and with probabilities that are all 0)
- c In which of the examples the set of outcomes is continuous?

2.7 Show with the aid of Venn-diagrams that the distributive properties are always valid:
$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$ and $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$

2.8 Two dice (6 faces) are thrown.
- a Which pairs of the following events are disjoint (mutually exclusive)?
  - A the total number of dots is even
  - B the total number of dots is odd
  - * C(i) the total number of dots is at least i (i = 9, 10, 11, 12) * this row may be omitted
  - D the total number of dots is a square
  - E the number of dots on one of the dice is equal to the number of dots on the other die.
- b Which of the above-mentioned events is a subset of another above-mentioned event?
- c Find the probability of each of the above-mentioned events assuming that the dice are unbiased. (There is a symmetrical sample space with 36 outcomes.)

2.9 A coin and a die are thrown. The probability that an outcome in the sample space of the coin, S_C, will occur, is of course 1. In the same way: $P(S_D) = 1$ (with S_D the sample space of the die). Someone reasons as follows: S_C and S_D are mutually exclusive, so it must be that: $P(S_C \cup S_D) = P(S_C) + P(S_D) = 1+1 = 2$. What is wrong in this argument?

2.10 A balanced coin is tossed five times. Find the following probabilities:
a exactly three heads.
b at least three heads.

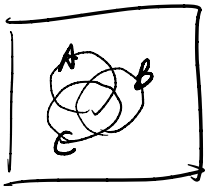
2.11 A few centuries ago the mathematician de Méré was confronted with a paradox. In games of chance in which three unbiased dice were thrown, he observed in an experimental way that a total number of 11 dots happened more often than a total number of 12 dots. According to his opinion this couldn't be true if you viewed the experiment in a theoretical way, because there are six possibilities of throwing a total number of 11 dots and six possibilities of throwing a total number of 12 dots (write down these possibilities yourself)! What is wrong in his reasoning and what are the exact probabilities?

2.12 An unbiased die is thrown three times. Find the following probabilities:
a exactly three times the outcome is even.
b exactly two times a six is thrown (and the other time no six appears).
c the total number of dots is at least seventeen.

2.13 You are given a die. Can you check whether this is an unbiased die? If yes, in which way, if not, why not?

2.14 Prove Theorem 2.3 in a formal way.

2.15 It is always true that:
$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$
a Prove this in a formal way (apply Theorem 2.6 repeatedly, starting with $P(A \cup (B \cup C))$.
b Show this with the aid of Venn-diagrams.



2.16 For a certain type of car the steering wheel and/or the brakes can be defective (other possible defects we leave out of consideration). Of all cars of this type 87% is all right, while 3% has only a faulty steering wheel, and 6% has only defective brakes . What percentage of cars has both defects?

2.17 Let A and B certain events. Explain why it is true that: $P(A \cap \overline{B}) \geq P(A) - P(B)$

2.18 Water may be contaminated with A or with B , or with both. Of the water samples is 20% pure, 40% is polluted with A and 50% is polluted with B. Calculate the probability that a sample contains exactly one of the two pollutions A or B.

2.19 Show that: $P(\overline{A} \cap B) = P(B) - P(A \cap B)$

2.20 Given is that $P(A) = P(B) = 1/4$ and $P(A \cap B) = 1/10$. Find $P(\overline{A} \cap B)$, $P(A \cup B)$ and $P(\overline{A} \cup \overline{B})$.

2.21 The event that *exactly one* of the two events A and B occurs will be written here as A ór B (which is thus something other than A or B; the notation XOR (exclusive or) is also used for this purpose). Give an expression for P(A ór B).

2.22 Prove the following equalities:
a $P(\overline{A} \mid B) = 1 - P(A \mid B)$
b $P(A_1 \cup A_2 \mid B) = P(A_1 \mid B) + P(A_2 \mid B) - P(A_1 \cap A_2 \mid B)$

2.23 Let us (for simplicity) assume here that the probability that a randomly selected person has a birthday in a given month is equal to 1/12 for each of the twelve months. Calculate the probability P_r that at least two persons from r randomly selected people have their birthday in the same month. Find this probability for r = 2 , 3, . . , 12.

2.24 Club AX wins with probability 0.7 each of its matches, regardless of the opponent and the previous results. Calculate the probability that AX over the next three games will win more often than not (either lose or draw).

2.25 The following problem is known as the 'Birthday problem' (Dutch: 'verjaardagsprobleem'). Assume that a year consists of 365 days (so leap years are not considered here) .
a Find the probability that at least 2 persons out of a group of 4 persons have the same birthday date.
b Find the probability that at least 2 persons out of a group of n persons have the same birthday date.

2.26 The following information is given about the three events A, B and C: A and B are independent; A and C are independent, B and C are disjoint (mutually exclusive). Moreover, the following probabilities are given:

$$P(\mathrm{B}) = 0.5 \quad P(\mathrm{C}) = 0.3 \quad P(\mathrm{A} \cup \mathrm{B} \cup \mathrm{C}) = 0.9$$

Find $P(\mathrm{A})$.

2.27 Student Caroline takes driving tests until she passes. Assume that all attempts are independent of each other and have a success rate of 4/7. Calculate the probability that Caroline needs an even number of attempts to pass the driving test.

2.28 A coin is tossed until the first time that 'heads' is tossed, ór till 'tails' have been thrown six times.
a Find the probability that the numbers of throws is even, assuming that the coin is unbiased (which means that at each toss both the probability of 'heads' and the probability of 'tails' is ½),
b The same question, but let us now assume that the probability of 'heads' is p.

2.29 Aircraft engines are known to operate independently from each other and they can show a defect with probability p during the flight. A flight will now be called successful if at least half of the aircraft engines do not show a defect during the flight. For which values of p do we have more chance of making a successful flight in a four-engined plane than in a twin-engined plane?

2.30 In an American metropolis three (daily) city newspapers are printed: I, II and III.
Suppose S = {all residents of the city}. Use the notation A (or B and C respectively) for the event that a random resident reads the newspaper I (or II and III respectively). Furthermore, suppose that:
$P(\mathrm{A}) = 0.10$, $P(\mathrm{B}) = 0.30$ and $P(\mathrm{C}) = 0.05$.
$P(\mathrm{A} \cap \mathrm{B}) = 0.08$, $P(\mathrm{A} \cap \mathrm{C}) = 0.02$, $P(\mathrm{B} \cap \mathrm{C}) = 0.04$ and $P(\mathrm{A} \cap \mathrm{B} \cap \mathrm{C}) = 0.01$.
Solve the following problems by using a Venn-diagram (shade every time the area of the event mentioned):
a What is the probability, if you randomly select an inhabitant from the city register, that this person reads daily **exactly one** of the three newspapers?
b What is the probability that this inhabitant reads **at least two** daily newspapers?
c Suppose I and III are morning papers and II is an evening paper. Find the probability that the randomly selected inhabitant daily reads **exactly one** morning paper **and** the evening paper.

2.31 A group of n people will draw lots for the Dutch Santa Claus 'surprises'. Everyone does a lot with his/her name on it in a box, and then everyone draws at random a lot. Once someone draws his/her own lot, the procedure must be repeated. What is the probability that no one draws his/her own lot? This problem is also known as the 'hat problem' (n men visit a pub, and throw their hat into the corner. When they are drunk and go home, they take at random a hat. How likely is it that no one takes his own hat?)
a Do you expect that this function increases or decreases as a function of n ? increase
b First, solve this problem for $n = 3$, and for $n = 4$.
c Consecutively, find a general formula for arbitrary values of n.

2.32 Both the events A and B have probability ½. Furthermore, it is given: $P(A \cup B) = {}^2/_3$.
a Are A and B mutually exclusive?
b Are A and B independent?
c Find $P(\overline{A} \cap B)$.
d Find $P(\overline{A} \cap \overline{B})$.
e Find the relation between $P(\overline{A} \cap \overline{B})$ and $P(A \cup B)$.

2.33 (B&E, 1.37) Let $P(\mathrm{A}) = 0.4$ and $P(\mathrm{A} \cup \mathrm{B}) = 0.6$.
a For what value(s) of $P(\mathrm{B})$ are A and B mutually exclusive?
b For what value(s) of $P(\mathrm{B})$ are A and B independent?

2.34 Of the events A, B and C is given: A and B are disjoint; A and C are independent; B and C are independent; $P(\mathrm{A}) = 0.3$; $P(\mathrm{B}) = 0.2$; $P(\mathrm{C}) = 0.4$. Find the probability that **exactly one** of the events A, B or C occurs.

2.35 A game at the fair goes like this: throw a coin with radius r on a large table on which are drawn adjacent squares (checker pattern) with side a ($a > 2r$) . If the coin, after the toss, does not cover any line on this table, then the player wins and receives a prize.
a What is the chance to win for a haphazardly throwing player?
b If r is equal to 1, which value(s) of a can be taken by the operator of this fairground attraction if he wants that the profit probability of a haphazardly throwing player is not greater than 0.4?

2.36 Prove that the independency of A and B implies that the complement of A is also independent of B. [From this it also follows that the complements of A and B are independent, can you see this?] Furthermore, prove the following statement: A and B are independent $\Leftrightarrow P(A \mid B) = P(A \mid \overline{B})$

In problems 2:37 up to and including 2:49 it can be useful (but not necessary) to draw a probability tree and/or to establish a cross table . Please practise!

2.37 Draw two cards from a deck of 52 cards without replacement. What is the probability that both cards are hearts, given that at least one of the two is a card of hearts?

2.38 Box I contains four green and five red marbles box II six green marbles and eight red marbles Someone draws one marble from box I and puts this marble in box II. Then, one marble is drawn from box II.
a Find the probability that the second marble drawn (the marble drawn from the "new" box II) is green.
b Find the probability that the first marble drawn is red knowing that the second one drawn was green.

2.39 (B&E, 1.33) One card is selected from a deck of 52 cards and placed in a second deck. A card then is selected from the second deck
a What is the probability the second card is an ace?
b If the first card is placed into a deck of 54 cards containing two jokers, then what is the probability that a card drawn from the second deck is an ace?
c Given that an ace was drawn from the second deck in (b), what is the conditional probability that an ace was transferred?

2.40 (B&E, 1.42) The probability that a marksman hits a target is 0.9 on any given shot, and the results of repeated shots are independent. He has two pistols; one contains two bullets and the other contains only one bullet. He selects a pistol at random and shoots at the target until the pistol is empty. What is the probability of hitting the target exactly once?

2.41 The firms A, B and C deliver 15, 25 resp. 60% of all memory sticks. The percentages of defective memory sticks are for the three firms respectively 5, 7 and 4 (%).
a Compute which percentage of the memory sticks is defective.
b Find the conditional probability that a memory stick was delivered by firm B, given that the memory stick is defective.

2.42 Twenty years ago, the tests on HIV infection were not as good as nowadays . When an infected person is tested, the probability that such a test proves to be positive is equal to 98%. If someone who has not been infected, is tested, the probability of a positive test result is equal to 1% (false positive). It is also known that within the population, 0.1% of all people is infected. A random person from the population is subjected to the HIV test, and that test proves to be positive. We are interested in the probability that the person is actually infected.
a Try first to estimate this probability by a "lucky guess".
b Now calculate this probability by using probability calculus.
c What happens to this probability if the person was not chosen at random from the population, but belongs to the so-called risk group?

2.43 Someone possesses 64 unbiased coins and one coin with 'heads' on both sides. From a box containing those 65 coins one coin is randomly selected. Without looking at this coin it is tossed six times. All six times 'heads' is tossed. Find the probability that there has been tossed with the coin with 'heads' on both sides.

2.44 In one room there are three cabinets. Each cabinet has two drawers. In one cabinet the first drawer contains a golden coin and the other drawer contains a silver coin. In the second cabinet both drawers contain a golden coin. In the last cabinet both drawers contain a silver coin. Now you walk into the room, not knowing what coins are where, and you open a random drawer. This appears to contain a golden coin. What is the probability that the other drawer of that cabinet contains a gold coin?

2.45 The **Monty Hall problem** (or Willem Ruis problem). Imagine you are the winner in a game show on TV, but the price has yet to be determined. The host shows you three doors, and says that there is a car behind

one of the doors, and a goat behind the other two doors. You pick a door but the door is not opened yet. The host (who knows where the car is located) opens one of the other two doors such that it shows a goat (if he has a choice of two doors with goats, he randomly chooses). The host is giving you the chance to change your initial choice. Is it advantageous to do so?

2.46 Have another look at the Monty Hall problem. Now imagine that the host, as he has the choice of two doors with goats, not chooses a random door to open, but systematically opens the door to the right of the door initially selected by you, unless of course the car is located there. (If you have already selected the rightmost door, the host will preferably enter the far left door). Is it at this host strategy now advantageous for you to change your initial choice?

2.47 The probabilities of failure in the application of methods A, B and C in a certain process are 30, 20 and 10% respectively. Method A is used twice as much as Method B, and four times as much as Method C.
- a Calculate the overall probability of failure in a single trial (in which either method A ór method B ór method C is used).
- b If in a certain situation the process fails, then find the probability that this happened while method B was applied.

2.48 A message is sent as a sequence of zeros and ones. The probability that a 0 will be sent at a given moment is 0.4; of course, otherwise a 1 is sent. A transmitted 1 will be, with probability 0.2, wrongly received as being a 0 and a transmitted 0 will with probability 0.1 erroneously be received as a 1.
- a Find the probability that at any arbitrary moment the first signal received will be a zero.
- b Find the probability that a 0 had been sent if it is given that a 1 is received.

2.49 You are told that a family, totally unknown to you, has two children, and that those two children are not *both* sons (so there is at least one daughter). Assume that the probability that a boy is born is equal to the probability that a girl is born.
- a Determine the probability that there are two daughters in the family.
- b Now you go to their home, you ring the bell, and a daughter opens the door. What is now the probability that there are two daughters in the family?

2.50 The following model is sometimes used to simulate the spread of a contagious disease. Start with a box of b black balls and r red balls. Draw from this box a random ball. Put it back together with c copies of the same colour. Then the process is repeated each time.
- a Calculate the probability that the first three balls drawn are red.
- b Show that the probability of a black ball at the first draw is equal to the unconditional probability of a black ball at the second draw.
- c Show that the probability of a black ball at the k-th draw is equal to the (unconditional) probability of a black ball at the first draw.

2.51 In a box of 25 old light bulbs five of them are defective, because of rough handling. They successively draw light bulbs out of this box. Find the probability that the second defective light bulb is found in the third draw if
- a each light bulb is replaced into the box after a check.
- b the bulbs are not replaced.

2.52 A small company that provides shuttle flights, is flying with planes with room for up to 8 people. The company has determined that the probability that a passenger (who had already bought a ticket) does not show up for the flight is 0.1. For each flight the company sells tickets to the first ten people who ordered a ticket. At the airport no tickets can be purchased. The probability distribution of the number of tickets sold per flight is as follows:

Number of tickets	6	7	8	9	10
Probability	0.3	0.3	0.25	0.1	0.05

Calculate the probability that the number of passengers showing up for a flight is higher than the number of available seats.

2.53 Difficult. A variant of the birthday-problem: Suppose there are 180 cyclists participating in the Tour de France, which lasts 23 days. What is the probability that two or more cyclists have their birthday on the same day somewhere during the duration of the Tour de France?

2.54 If A, B and C are connected in parallel with reliabilities p_A, p_B, p_C (probabilities independently of each other), then determine the reliability of the system as
- a at least one component has to function.
- b at least two components have to function.

2.55 A system consists of two sub-systems connected in series, each consisting of two components: the first sub-system contains A and B in parallel, and the second one C and D in parallel. Determine the reliability of the entire system, if both A and B function with probability p_A , and both C and D function with probability p_C (probabilities independently of each other).

2.56 The digits 1, 2 , 3, . . . , 9 are put one after the other in any arbitrary order, resulting in a number of nine digits. Find the probability
- a that the resulting number is even.
- b that the resulting number can be divided by 5.
- c that in the resulting number the 4 and 6 are adjacent.

2.57 *n* people, including A and B, are put in a row at random. Calculate the probability that A will *not* be next to B.

2.58 From a group of 40 men and 16 women a subgroup of eight people will be drawn at random. Calculate the probability that this group consists only of men.

2.59 In a particular game a player has to choose ten different numbers from { 1, 2, . . . 80 }. The casino draws 20 different numbers from the 80 available. Find the probability that the player chose exactly five numbers correctly (so five of the 10 numbers chosen will appear in the 20 numbers drawn by the casino).

2.60 Annabel, Bert, Carice, Daan and Eline are friends. We put three of them in a row.
- a Write down (with the aid of their initials) all possible permutations.
- b Write down all possible combinations. How many permutations lead to one and the same combination?

2.61 A bowl contains eight red and five yellow marbles. Three marbles are randomly drawn from this bowl.
- a1 Find the probability that no yellow marble is drawn, if the three marbles are drawn without replacement.
- a2 The same question, but now the three marbles are drawn with replacement.
- b The same questions as in part a, but now the bowl contains 24 red and 15 yellow marbles.
- c What happens if the number of marbles in the pot is increased while making sure that the ratio between the number of red and yellow marbles remains unchanged (so equal to 8/5)?

2.62
- a Take without replacement at random three students out of a class of 20 students. Then, select randomly again three students out of the entire class of 20 students. Calculate the probability that the two triplets have at least one student in common.
- b Draw twice a “bridge hand” (13 playing cards) out of a complete deck with 52 cards (the two drawings are from different decks of cards). Find the probability that both hands have at least one card in common.

2.63 Find the probability that an arbitrary “bridge hand” (13 playing cards out of a game of 52)
- a contains no card of hearts.
- b contains exactly two aces.

2.64 A pond contains 50 fish, 10 of which are marked. Someone captures at random seven fish. Find the probability that exactly two marked fish were among the captured fish.

2.65 Suppose that 50 states delegate each two senators to a meeting. At random 50 senators are elected at this meeting for a committee. Calculate the probability that a given state X is represented in this committee.

2.66 A large company is housed in a building with eleven floors, one ground-floor and ten upper storey floors, numbered from 1 to 10. Seven employees of the company step into the elevator on the ground floor. Suppose they decide independently of each other on which floor they go off and that each floor has an equal probability of being chosen. What is the probability that the elevator on his way up will have to stop exactly seven times before everyone got off? What is the probability that the elevator will have to stop exactly six times?

2.67 The showroom of a car dealer contains 20 cars, numbered from 1 to 20 depending on the net cost price of the vehicle. Number 1 stands for the cheapest car, for example, number 19 for the second most expensive one. The owner keeps the car keys in a big box in his office. Suppose that the son of the owner is blindfolded on his eighteenth birthday and he may select at random three car keys out of the box. From these three keys he can then choose a car as a birthday present. What is the probability that all three of the car keys carry a number that is greater than or equal to 17? Find also the probability that at least one of the car keys carries a number that is greater than or equal to 17.

2.68 In the attic of the house of a war veteran from World War II is a box with 18 equally large oblong pieces of fabric: six red, six white and six blue. One evening the veteran submits to his wife (a lady who once followed an evening course in probability calculus) the following problem: "Suppose I haphazardly would pull 5 pieces of fabric out of the box. What is the probability that I can sew the Dutch flag with these 5 pieces?
After fifteen minutes the woman comes up with the following reasoning: " It is clearly the probability that all three colors are represented at least once. This probability is equal to
1 – *P*(two colours are missing) - *P*(one colour is missing).
Furthermore, we know $\boldsymbol{P}(\text{two colours are missing}) = \binom{3}{2} \times \binom{6}{5} \Big/ \binom{18}{5} = 0.0021$
Moreover, $P(\text{one colour is missing}) = \binom{3}{1} \times \binom{12}{5} \Big/ \binom{18}{5} = 0.2773$
This means that your requested probability is equal to $1 - 0.0021 - 0.2773 = 0.7206$."
Is her way of reasoning correct?

2.69 Prove Theorem 2.10.

2.70 Prove that $\binom{n}{k} = \frac{n-k+1}{k}\binom{n}{k-1}$.

2.71 Prove Theorem 2.11.

2.72 In how many different ways the nine letters of the (Dutch) word *repetitie* can be arranged?

2.73 Three boys and three girls randomly take place in a row of chairs. What is the probability that boys and girls always alternate in the row (never two boys or two girls adjacent to each other) ?

2.74 License plate numbers consist of two digits, followed by three letters, then one digit again .
a How many different license plate numbers are possible, if all the letters of the alphabet are allowed, and if letters and numbers can be used repeatedly?
b How many different license plate numbers are possible, if letters cannot be used repeatedly?

2.75 From 18 students six have level A, four have level B, and the rest level C.
In how many ways can these 18 levels be distributed among the 18 students?

2.76 Consider the 12 letters of the word HIPPOPOTAMUS.
Find the probability that at an arbitrary ranking of these letters the letter "H" will be placed next to a letter "O" (hint: Consider two possibilities for the "H" position, namely somewhere in the middle of the word, or otherwise somewhere at one of its ends (either first or last position)).

2.77 (B&E, 1.68) Suppose 14 students have tickets for a concert. Three students (Bob, Jim and Tom) own cars and will provide transportation to the concert. Bob's car has room for three passengers (nondrivers), while the cars owned by Jim and Tom each has room for four passengers.
a In how many different ways can the 11 passengers be loaded into the cars?
b At the concert hall the students are seated together in one row. If they take their seats in random order, find the probability that the three students who drove their cars have adjoining seats.

2.78 When you are playing poker with dice, five dice are rolled independently from each other.
The following situations may then arise:
– **Five of a kind**: 1 score appears five times.
– **Four of a kind**: 2 different scores appear, 1 score shows up four times and the other score once.
– **Full house**: 2 different scores appear, 1 score shows up three times and the other score twice.
– **Three of a kind**: 3 different scores appear, 1 score shows up three times and each of the other 2 one time.
– **Two pair**: 3 different scores appear, 2 scores show up twice and the third score once.
– **One pair**: 4 different scores appear, 1 score shows up twice and each of the other 3 scores one time.
– **None alike**: 5 different scores appear.
Find the probability of all of these situations, which are also called "poker dice hands" .

3 Discrete random variables

3.1 Introduction

A **random variable** (Dutch: kansvariabele, stochastische variabele, stochast) can be seen as a function that assigns a real number to each possible outcome in a sample space *S*. Random variables are usually represented by capitals, often *X* or *Y*. To emphasize that a random variable is actually a function, we could use the following notation: $X(e) = x$, where *e* any outcome such that $e \in S$. Lowercase letters like *x* or *y* are used to denote possible values that the corresponding random variables *X* and *Y* can take on. Usually, we will use the shorter notation *X* instead of $X(e)$ as can be illustrated in the example below:

Example 3.1

Experiment: toss a coin, and define the sample space as *S* = { H(ead), T(ail) }. We can, for example, define the following random variable *X*:

$$X(e) = \begin{cases} 0, & \text{for } e \in \{\text{H}\} \\ 1, & \text{for } e \in \{\text{T}\} \end{cases}$$

This will be abbreviated as:

$$X = \begin{cases} 0, & \text{if H(ead)} \\ 1, & \text{if T(ail)} \end{cases}$$

With this definition, the probability that *X* attains the value 0 is the same as the probability that a toss of the coin results in a Head. Using the probability set function $P(\cdot)$ (see Axiom's, section 2.4.4), we should formally denote this as:

$$P(\{e \mid e \in S \text{ and } X(e) = 0\}) = P(\{H\}) .$$

The left-hand side is unnecessarily complicated, and we will write instead: $P(X = 0)$.

If the coin is unbiased, then $P(X = 0) = \frac{1}{2}$ and $P(X = 1) = \frac{1}{2}$, or $P(X = x) = \frac{1}{2}$ for $x = 0, 1$. ◄

Remark: purists might object that using the notation $P(X = 0)$, we use a probability set function defined in a different way, because the sample space has changed from the original space into the image of the original one by the transformation as defined by the function *X*. That is the reason why B&E uses a different type of brackets $P[\cdot]$ instead of $P(\cdot)$. But here, we will use no such distinction in notation.

A random variable is called *discrete*, when its set of possible outcomes is a specified finite or countable list of values $\{x_1, x_2, \ldots\}$, or *continuous* if it can take on any (real) value in an interval or collection of intervals (e.g. $\{x \in \mathbb{R} \mid 1 < x \leq 3\}$. Mixtures of both types are also possible.

Example 3.2

Experiment: throw two dice. If *X* is the total number of dots, then *X* is clearly discrete.
Experiment: measure the time *X* (in seconds) until the next radioactive decay occurs in a piece of material. If *X* is measured with infinite precision, then *X* is a continuous random variable. But if the outcome is rounded to the nearest integer value, the set of outcomes is countably infinite, and *X* is (strictly speaking) discrete. ◄

This chapter will deal with discrete random variables. We will see later, in the next chapter, that we will need different techniques for continuous random variables, although many results will still hold for both types.

3.2 Probability distribution functions

(B&E, page 56-62)

A **probability distribution** is characterised by the set of possible outcomes for a random variable along with all the possible probabilities for each subset of those outcomes. A probability distribution for a discrete random variable can be specified by giving its probability distribution function, its

cumulative distribution function, or just by naming its type with the associated values for the parameters (as we will see in section 3.7). First, we define the probability distribution function:

Definition 3.1 *(B&E, Def. 2.2.1)*

If the set of all possible values of a random variable X is the countable set $\{x_1, x_2, \ldots\}$, then X is called a **discrete random variable**. The **probability distribution function (pdf)** $f(x)$ (or $f_X(x)$) for a discrete random variable X is defined for any real x as:

$$f(x) = \begin{cases} P(X = x) & \text{if } x \in \{x_1, x_2, \ldots\} \\ 0 & \text{if } x \notin \{x_1, x_2, \ldots\} \end{cases}$$

(Dutch: kansfunctie). Alternatively, this function is often called a probability mass function.

Remark: Some books, like B&E use the term probability *density* function; this is somewhat confusing since in the discrete case the assigned values are no 'densities' in the usual sense. B&E also defines the pdf only for $x \in \{x_1, x_2, \ldots\}$, which is unnecessarily restrictive and leads to some slightly incorrect results later on.

The set $\{x_1, x_2, \ldots\}$ is also called the **support** (Dutch: drager) of $f(x)$. According to the definition above, the pdf assigns the value 0 to all real values x which are not part of the support set. Whenever we specify a pdf in this reader, we will always have to specify the support along with the probabilities for values of x which are part of the support. Implicitly, this means that the pdf will be 0 for all other values of x.
It is very straightforward to see that each discrete pdf should have the following two properties:

$$f(x) \geq 0 \quad (\text{for } x \in \mathbb{R}) \qquad \text{and} \qquad \sum_{\text{all } x} f(x) = 1$$

(For simplicity of notation, we will use in summations often 'all x', short for $x \in \{x_1, x_2, \ldots\}$.)

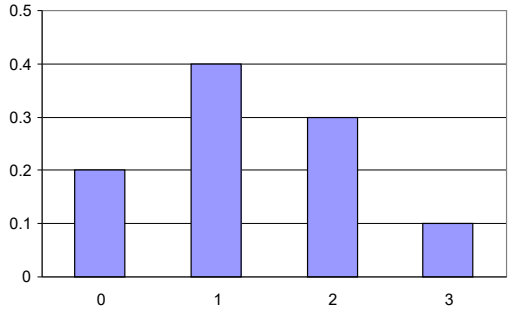
Often it is possible to specify a pdf in the form of a formula, as we will see frequently. But sometimes it is more convenient to give the pdf in the form of a table. To illustrate a pdf graphically, often bar charts are used, or a chart with just vertical lines, instead of columns.

Example 3.3

Sales data show that the number of cars sold per day by the Best Cars dealership is distributed as follows: no car is sold on 20% of all days, one single car is sold at 40% of all days, two cars at 30% and three cars at 10% of all days. Define X as the number of passenger cars sold on any arbitrary day. We can create the following table for the probability distribution function of X:

x	0	1	2	3
$f(x)$	0.2	0.4	0.3	0.1

Graphically:



Check for yourself that each of the two properties for a discrete pdf is satisfied. ◄

Definition 3.2 *(B&E, Def. 2.2.2)*

The **cumulative distribution function** (**CDF**) of a random variable X is defined for any real x by:

$$F(x) = P(X \le x)$$

(Dutch: verdelingsfunctie).

Note that the very common convention of using either small letters or capitals tells us already if we are dealing with a pdf (f), or with a CDF (F). The CDF of a discrete random variable can be deduced directly from the pdf by summation of probabilities:

$$F(x) = P(X \le x) = \sum_{\text{all } u \le x} f(u) .$$

Vice versa, it is also possible to derive the pdf from a given CDF. Formally:

$$f(x) = P(X = x) = P(X \le x) - \lim_{h \downarrow 0} P(X \le x - h) = F(x) - \lim_{h \downarrow 0} F(x - h) .$$

Example 3.4

We will determine the CDF of the random variable in Example 3.3. For example, we get

$$F(1) = P(X \le 1) = f(0) + f(1) = 0.2 + 0.4 = 0.6 .$$

In fact, we will get the same answer for any value of $1 \le x < 2$, e.g. $F(1.9999) = 0.6$. (check!).

But at $x = 2$, we observe a jump in the function value:

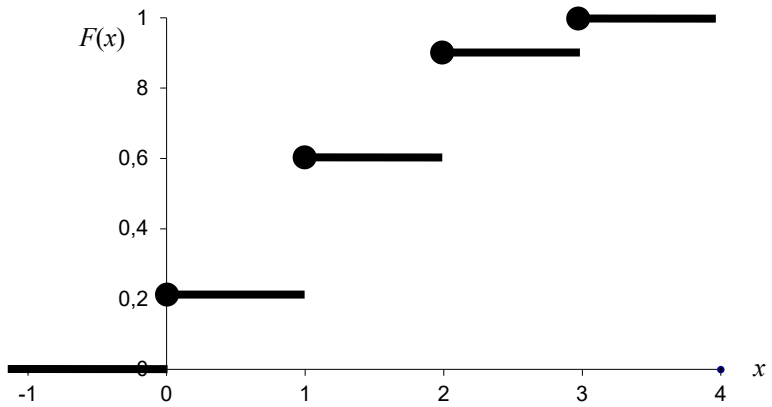
$$F(2) = P(X \le 2) = f(0) + f(1) + f(2) = 0.9 \text{ !}$$

In this way, we can specify the following table for the CDF:

	$x < 0$	$0 \le x < 1$	$1 \le x < 2$	$2 \le x < 3$	$3 \le x$
$F(x)$	0	0.2	0.6	0.9	1

Graphically, the CDF looks like:

We see in the figure above the specific shape for the CDF of a discrete random variable, a so-called



step-function. The sizes of the steps or jumps are equal to the values of the pdf at those points. In case we need to derive the pdf for a given CDF, we only need to look at those points where the jumps occur. The collection of those points is the support set, and the values for the pdf at those points can be found by determining the size of the jumps. In this example, we see a jump occurring at $x = 0$, also seen from the table and because $F(0) = P(X \le 0) = 0.2$ is greater than $F(-0.000001) = 0.0$. So $f(0) = F(0) - \lim_{h \downarrow 0} F(0 - h) = 0.2 - 0.0 = 0.2$. Similarly, we get

$f(1) = 0.6 - 0.2 = 0.4$. Note that the formula $f(x) = F(x) - \lim_{h \downarrow 0} F(x-h)$ also gives the value 0 for $x \notin \{x_1, x_2, \ldots\}$, e.g. $f(1.7) = F(1.7) - \lim_{h \downarrow 0} F(1.7-h) = 0.6 - 0.6 = 0.0$. ◄

It shall be clear now that any CDF is a nondecreasing function (i.e. $F(a) \le F(b)$ whenever $a < b$) and that $F(-\infty) = 0$ and $F(\infty) = 1$. We can also show that the CDF is always continuous from the right (i.e. $\lim_{h \downarrow 0} F(x+h) = F(x)$ (but not continuous from the left, because otherwise no jumps would be possible; in that case we are dealing with a continuous random variable, see next Chapter).

In section 3.7, we will encounter discrete random variables with a distribution which is of a specific type, like for example the binomial distribution. The binomial distribution is characterised (see later) by two parameters, *n* and *p*. In those cases, it is not necessary to specify a formula or table for the pdf or CDF, but it is sufficient to write $X \sim \text{BIN}(n, p)$, which can be read as 'X has a binomial distribution with parameters *n* and *p*'.

3.3 Expected value, variance, standard deviation and other measures

(B&E, page 61, 71-74)

Definition 3.3 *(B&E, Def. 2.2.3)*

The **expected value** or expectation or mean (Dutch: verwachte waarde of verwachting), notation: $\text{E}(X)$ or $\text{E}X$ or μ), of a discrete random variable is defined by:

$$\mu = \text{E}(X) = \sum_{\text{all } x} x\, f(x)$$

If the support set is infinitely large, it is possible that the sum $\sum_{\text{all } x} |x|\, f(x)$ does not converge (not 'absolutely convergent', and we say that the expected value does not exist.
We can also say that the expected value is equal to the weighted average of all possible values of *X,* where the probabilities for those values are used as the weights. Note that the symbol μ is used as well in the definition; in Chapter 1 this symbol has been used for the population mean. Here it is not different: μ represents again the population mean. The population is the (possibly infinite) set of all observations on the random variable *X* that can be performed. Within that population a proportion $f(x_1)$ of all observations has the value x_1, a proportion $f(x_2)$ of all observations has the value x_2, etc., which explains why the population mean can be found using the above formula.

Example 3.5

Continuation of Example 3.3. The expected value of *X* is

$$\text{E}(X) = \sum_{\text{all } x} x\, f(x) = 0 \cdot 0.2 + 1 \cdot 0.4 + 2 \cdot 0.3 + 3 \cdot 0.1 = 1.3$$

which is indeed equal to the mean number of cars sold per day (= μ). ◄

We will often encounter functions (transformations) of a random variable *X*. Say $g(\cdot)$ is a function which assigns to each possible value *x* of *X* the value $g(x)$ We can then write: $Y = g(X)$, where *Y* represents again a random variable (recall that a random variable is just a function or rule which assigns a numerical value to each possible outcome in a sample space). So *Y* will also have a pdf $f_Y(y)$, which can be derived from the pdf $f_X(x)$ of *X*.

Example 3.6

Continuation of Example 3.3. Say the profit of Best Cars for each car sold is € 2,000, minus the fixed costs of € 2,300 per day. Define the random variable *Y* as the profit on an arbitrary day. *Y* can then be seen as a function of *X*:

$Y = 2000X - 2300$.

The values Y can take are -2300, -300, 1700 and 3700 (follows directly from the values of X). The pdf $f_Y(y)$ can thus be simply represented by the following table:

y	-2300	-300	1700	3700
$f_Y(y)$	0.2	0.4	0.3	0.1

The expected value of Y is $\mathrm{E}(Y) = -2300 \cdot 0.2 - 300 \cdot 0.4 + 1700 \cdot 0.3 + 3700 \cdot 0.1 = 300$ (€) ◄

As we have seen in the example above, $\mathrm{E}(Y)$ can be found by first determining the pdf of Y. However, this detour is not necessary in view of the following important theorem (given here without proof, because the formal proof is still surprisingly complicated):

Theorem 3.1 *(B&E, Th. 2.4.1)*

If X is a discrete random variable with pdf $f(x)$, and $g(\cdot)$ is a real valued function on the support set of X, then:

$$\mathrm{E}(g(X)) = \sum_{\text{alle x}} g(x)\, f(x)$$

The following theorem can be very helpful when evaluating expected values.

Theorem 3.2 *(B&E, Th. 2.4.2)*

If X is a random variable with pdf $f(x)$ and $g(\cdot)$, $g_1(\cdot)$, $g_2(\cdot)$ are real functions (on a domain including the support of X) and a, b, and c are constants, then:

1. $\mathrm{E}(c) = c$ for any constant c
2. $\mathrm{E}(cg(X)) = c\mathrm{E}(g(X))$ for any constant c,
3. $\mathrm{E}(g_1(X) + g_2(X)) = \mathrm{E}(g_1(X)) + \mathrm{E}(g_2(X))$
4. $\mathrm{E}(aX + b) = a\mathrm{E}(X) + b$ for any constants a and b

Proof

1. $\mathrm{E}(c) = \sum_{\text{all } x} c \cdot f(x) = c \sum_{\text{all } x} f(x) = c$
2. $\mathrm{E}(cg(X)) = \sum_{\text{alle } x} cg(x) f(x) = c \sum_{\text{alle } x} g(x) f(x) = c\mathrm{E}(g(X))$
3. $\mathrm{E}(g_1(X) + g_2(X)) = \sum_{\text{alle } x} (g_1(X) + g_2(X)) f(x)$
   $= \sum_{\text{alle } x} g_1(X) f(x) + \sum_{\text{alle } x} g_2(X) f(x) = \mathrm{E}(g_1(X)) + \mathrm{E}(g_2(X))$
4. Try yourself; first apply rule 3, followed by rules 2 and 1.

Example 3.7

Continuation of Example 3.6. The expected value of the profit Y could also have been found by:

$\mathrm{E}(Y) = \mathrm{E}(2000X - 2300) = 2000 \cdot \mathrm{E}(X) - 2300 = 2000 \cdot 1.3 - 2300 = 300$ (€) ◄

The expected values of some functions of X are very important:

Definition 3.4 *(B&E, Def. 2.4.1)*

The **variance** (Dutch: variantie), notated by $\mathrm{Var}(X)$ or σ^2 of a discrete random variable X is defined by:

$$\sigma^2 = \mathrm{Var}(X) = \mathrm{E}((X-\mu)^2) = \sum_{\text{all x}} (x-\mu)^2 \, f(x)$$

The **standard deviation** (Dutch: standaardafwijking, standaarddeviatie) σ is the square root of the variance.

From the formula, we can see that the variance is a weighted average of the squared distances between each of the possible values of X and the mean, with the probabilities as weights. Distributions where the individual outcomes lie often far from the mean will, in general, have a large variance, and vice versa. The variance is a measure for the spread of a random variable.

The variance can also be written differently:

Theorem 3.3 *(B&E, Th. 2.4.3)*

If X is a random variable, then:

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - \mu^2 \quad (= \mathrm{E}(X^2) - (\mathrm{E}(X))^2)$$

Proof

$$\begin{aligned}\mathrm{Var}(X) &= \mathrm{E}((X-\mu)^2) \\ &= \mathrm{E}(X^2 - 2\mu X + \mu^2) \\ &= \mathrm{E}(X^2) - 2\mu\mathrm{E}(X) + \mu^2 \\ &= \mathrm{E}(X^2) - \mu^2 \qquad (\text{because } \mathrm{E}(X) = \mu)\end{aligned}$$

Note that the theorem above leads directly to the equation $\mathrm{E}(X^2) = \sigma^2 + \mu^2$.

Theorem 3.4 *(B&E, Th. 2.4.4)*

If X is a random variable, and a and b are constants, then:

$$\mathrm{Var}(aX+b) = a^2\mathrm{Var}(X).$$

Proof

$$\begin{aligned}\mathrm{Var}(aX+b) &= \mathrm{E}((aX+b-a\mu_X-b)^2) \qquad \text{(follows from Definition 3.4 and Theorem 3.2)} \\ &= \mathrm{E}((aX-a\mu_X)^2) \\ &= \mathrm{E}(a^2(X-\mu_X)^2) \\ &= a^2\mathrm{E}((X-\mu_X)^2) \qquad \text{(follows from Theorem 3.2)} \\ &= a^2\mathrm{Var}(X)\end{aligned}$$

(The constant b results in a shift of the distribution over a distance of b units. But such a shift does not change the spread of the distribution at all, so it should be no surprise that the value of b is irrelevant for the variance.)

Definition 3.5 *(B&E, Def. 2.4.2)*

The k-th **moment** of a random variable X is defined by:

$$\mathrm{E}(X^k)$$

The k-the **central moment** of a random variable X is defined by:

$$\mathrm{E}((X-\mu)^k)$$

Note that the expected value of X is the same as the 1st moment, and the variance is the same as the 2nd central moment. Also note that the first central moment is always equal to 0. The third central moment is a measure of the asymmetry of a distribution, the skewness (See p. 6). If a distribution is skewed to the right, the third central moment is positive (and the other way around).

3.4 The inequalities of Markov, Chebyshev and Jensen

(B&E, 75-76)

The inequalities discussed in this paragraph are important from a theoretical point of view.

Theorem 3.5 (Markov's inequality) *(≈B&E, Th. 2.4.6)*

If X is a random variable and $c > 0$, then:

$$P(|X| \geq c) \leq \frac{\mathrm{E}(|X|)}{c}$$

Proof

Define S as the sample space of X. We will split S into two subsets:

A_1 is the subset with outcomes x such that $|x| \geq c$

A_2 is the subset with outcomes x such that $|x| < c$

$$\begin{aligned}
\mathrm{E}(|X|) &= \sum_{x \in S} |x| f(x) \quad \text{(Theorem 3.1)} \\
&= \sum_{x \in A_1} |x| f(x) + \sum_{x \in A_2} |x| f(x) \\
&\geq \sum_{x \in A_1} |x| f(x) \quad \text{(second term above } \geq 0, \text{ since both } |x| \geq 0 \text{ and } f(x) \geq 0) \\
&\geq \sum_{x \in A_1} c\, f(x) \quad \text{(because for each } x \text{ in } A_1 \text{ we know that } |x| \geq c) \\
&= c\, P(X \in A_1) \\
&= c\, P(|X| \geq c)
\end{aligned}$$

Remark: many different versions exist for this inequality, all called Markov's inequality.

In the course Probability Theory and Statistics 3, the inequality of Chebyshev will be used in the proof of several important theorems (like the Law of Large Numbers). It states that the probability that a r.v. will deviate at least k standard deviations from its mean is at most equal to $1/k^2$. The proof is similar to the proof above (in exercise 3.18, an alternative proof will be asked based on Markov's inequality).

Theorem 3.6 (Chebyshev's inequality) *(B&E, Th. 2.4.7)*

If X is a random variable with expected value μ and standard deviation σ, then for any $k > 1$:

$$P(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2}$$

Proof

Define S as the sample space of X. We will split S in two subsets:
S_1 is the subset with outcomes x with a distance from μ less than $k\sigma$, so $|x-\mu|<k\sigma$
S_2 is the subset with outcomes x with a larger distance from μ, so $|x-\mu|\geq k\sigma$ such that $|x|<c$

$$\sigma^2 = \text{Var}(X) = \sum_{x\in S}(x-\mu)^2 f(x) \quad \text{(Definition 3.4)}$$

$$= \sum_{x\in S_1}(x-\mu)^2 f(x) + \sum_{x\in S_2}(x-\mu)^2 f(x)$$

$$\geq \sum_{x\in S_2}(x-\mu)^2 f(x) \quad \text{(because } (x-\mu)^2 \geq 0\text{)}$$

$$\geq \sum_{x\in S_2}(k\sigma)^2 f(x) \quad \text{(because for each } x \text{ in } S_2 \text{ we know that } (x-\mu)^2 \geq (k\sigma)^2\text{)}$$

$$\Rightarrow \sigma^2 \geq (k\sigma)^2 \sum_{x\in S_2} f(x)$$

$$\Rightarrow \frac{1}{k^2} \geq \sum_{x\in S_2} f(x)$$

The sum on the right-hand side is exactly the probability of X attaining a value in S_2, which is the probability $P(|X-\mu|\geq k\sigma)$. Therefore:

$$\frac{1}{k^2} \geq P(|X-\mu|\geq k\sigma)$$

From the theorem we can immediately see that $P(|X-\mu|<k\sigma) \geq 1-\frac{1}{k^2}$ (for $k>1$). Note that this last inequality can also be written as: $P(\mu-k\sigma<X<\mu+k\sigma)\geq 1-\frac{1}{k^2}$.

When we, for example, set $k=3$, then this inequality tells us that the probability of X assuming a value which is within three standard deviations of the mean of X is at least 88.9% (=8/9).

In Chapter 1, we discussed how we can interpret the standard deviation if the distribution of a data set displays a typical bell shape. If a certain probability distribution has this bell shape, then we could say, for example, that the probability of X assuming a value which is within three standard deviations of the mean is greater than 0.99. When we compare this with the result above using Chebyshev's inequality (>0.889), we see that Chebyshev's inequality seems much less precise. The reason is simple: Chebyshev's inequality applies *always*, so even if the distribution has a very different shape than the bell shape. It also stresses the point that it is usually not very useful to try to find probabilities using this inequality, since it will, in general, result in very rough estimates only.

Another important inequality is Jensen's inequality, which is *not* discussed in Bain & Engelhardt.

Theorem 3.7 (Jensen's inequality)

If X is a random variable, $g(\cdot)$ is a convex function (on an interval which includes the complete support set of X) and both E(X) and E(g(X)) exist, then:

$$g(\text{E}(X)) \leq \text{E}(g(X))$$

If $g(\cdot)$ is strictly convex, then equality only occurs when $P(X=c)=1$ where $c=\text{E}(X)$.

Proof

Since $g(\cdot)$ is convex, there exists for each given value a a value b, such that $g(x)\geq g(a)+b(x-a)$ for all x (informally: a convex function always lies above the tangent at

$x = a$). In case that $a = \mathrm{E}[X]$, we get $g(x) \geq g(\mathrm{E}(X)) + b(x - \mathrm{E}(X))$ for some b. When we take the expected value of the functions on both sides of this inequality, we find:

$$\mathrm{E}(g(X)) \geq \mathrm{E}(g(\mathrm{E}(X))) + \mathrm{E}(b(x - \mathrm{E}(X))) = \mathrm{E}(g(\mathrm{E}(X))) + 0 = g(\mathrm{E}(X)) .$$

If $g(\cdot)$ is strictly convex, then the strict inequality $g(x) > g(\mathrm{E}(X)) + b(x - \mathrm{E}(X))$ is satisfied for any $x \neq \mathrm{E}(X)$, so equality of $g(\mathrm{E}(X))$ and $\mathrm{E}(g(X))$ can occur only when $P(X = \mathrm{E}(X)) = 1$.

Corollary: If $g(\cdot)$ is concave, then $-g(\cdot)$ is convex, and the previous theorem applies with the result that $-g(\mathrm{E}(X)) \leq \mathrm{E}(-g(X))$. This last result is equivalent to $g(\mathrm{E}(X)) \geq \mathrm{E}(g(X))$.

Example 3.8

1. The function $g(x) = x^2$ is strictly convex on the real line. If both expected values exist, we can apply Theorem 3.7, and thus: $(\mathrm{E}(X))^2 \leq \mathrm{E}(X^2)$, or $\mu^2 \leq \mathrm{E}(X^2)$. Check that equality will occur when X can attain only one single value (which will then necessarily be equal to $\mathrm{E}(X)$)
2. If X is a random variable with only positive outcomes, then $\mathrm{E}(\log(X)) \leq \log(\mathrm{E}(X))$, because the logarithm is a concave function. ◄

3.5 The probability generating function

Probability generating functions are particularly useful for discrete random variables which only have the natural numbers (0, 1, 2, ...) as the support set. For more general situations (continuous random variables, and/or negative outcomes), we prefer to use *moment generating functions* (see next chapter).

Definition 3.6

The **probability generating function** (Dutch: kansgenererende functie, kgf) of a random variable X which has only nonnegative integers as possible outcomes, is defined by:

$$G_X(t) = \mathrm{E}(t^X) = \sum_{x=0}^{\infty} t^x \; P(X = x)$$

The idea is simple: write a polynomial (power series) with all possible powers of a variable *t,* where the coefficient for t^x is the probability of $X = x$. Note that this is the same as the expected value of the function t^X.

Example 3.9

The probability generating function for the random variable X as defined in Example 3.3 is:

$$G_X(t) = 0.2 + 0.4t + 0.3t^2 + 0.1t^3$$ ◄

Quite often, we will be able to rewrite the power series in a much more convenient functional form. For example, as will be shown in section 3.7, the probability generating function of binomially distributed random variables can be written as $G_X(t) = (pt + q)^n$.

When we replace t by 0 in the probability generating function, we see directly that $G_X(0) = P(X = 0)$. But each probability $P(X = i)$ for i = 1, 2, … can also be derived from $G_X(t)$, which explains the term 'probability generating'. If we want to determine $P(X = i)$ while $G_X(t)$ is written in the form of the power series in the definition, then it is very simple: just take the coefficient associated with t^i. If $G_X(t)$ is given in another functional form and the corresponding power series cannot be found in a simple way, then we can still find those probabilities by taking (repeatedly) derivatives with respect to

t. To obtain $P(X = 1)$, we take the first derivative with respect to t, and then enter $t = 0$. This can be seen as follows:

$$G_X'(t) = \frac{d}{dt}G_X(t) = \frac{d}{dt}\sum_{x=0}^{\infty} t^x f(x) = \sum_{x=0}^{\infty}\frac{d}{dt}t^x f(x) = \sum_{x=1}^{\infty} x t^{x-1} f(x) = 1!\,f(1) + \sum_{x=2}^{\infty} x t^{x-1} f(x)$$

$$\Rightarrow \quad G_X'(0) = f(1)$$

For larger values of *i,* we can find the general result: $P(X = i) = G_X^{(i)}(0)/i!$ (see exercise 3.25).

Example 3.10

Consider the probability generating function $G_X(t) = e^{4(t-1)}$.

We can determine $P(X = i)$ for $i = 0$, 1 and 2 as follows.

$P(X = 0) = G_X(0) = e^{-4}$.

$G_X'(t) = 4e^{4(t-1)} \quad \Rightarrow \quad P(X = 1) = G_X'(0) = 4e^{-4}$

$G_X''(t) = 16e^{4(t-1)} \quad \Rightarrow \quad P(X = 2) = G_X''(0)/2! = 8e^{-4}$ ◄

Instead of entering $t = 0$, we can also replace t by 1, giving the following results:

Theorem 3.8 *(B&E, Th. 2.5.4)*

If X is a nonnegative, integer valued random variable with probability generating function $G_X(t)$, then:

$G_X(1) = 1$

$G_X'(1) = \mathrm{E}(X)$

$G_X''(1) = \mathrm{E}(X(X-1))$

Proof

$$G_X(1) = \sum_{x=0}^{\infty} 1^x\, P(X = x) = \sum_{x=0}^{\infty} P(X = x) = 1$$

We already know that $G_X'(t) = \sum_{x=1}^{\infty} x t^{x-1} f(x) = \sum_{x=0}^{\infty} x t^{x-1} f(x)$

$$\Rightarrow \quad G_X'(1) = \sum_{x=0}^{\infty} x f(x) = \mathrm{E}(X)$$

and $G_X''(t) = \sum_{x=0}^{\infty} x(x-1) t^{x-2} f(x)$

$$\Rightarrow \quad G_X''(1) = \sum_{x=0}^{\infty} x(x-1) f(x) = \mathrm{E}(X(X-1)) = \mathrm{E}(X^2) - \mathrm{E}(X)$$

Remark: $\mathrm{E}(X(X-1))$ is also called the second *factorial* moment, which explains the alternative name 'factorial moment generating function' used by some authors instead of 'probability generating function'.

With the help of the theorem above, we can determine the expected value and / or the variance.

Example 3.11

Continuation of Example 3.10, where $G_X(t) = e^{4(t-1)}$. When we need to determine $\mathrm{E}(X)$ and $\mathrm{Var}(X)$, then we could try to rewrite this probability generating function as a power series, and use the coefficients to find the pdf of *X.* But a much more simple way is by applying the previous theorem.

$G_X'(t) = 4e^{4(t-1)} \quad \Rightarrow \quad \mu = \mathrm{E}(X) = G_X'(1) = 4$

To find the variance, we recall first that $\text{Var}(X) = \text{E}(X^2) - \mu^2$ which in turn can be written as $\text{Var}(X) = \text{E}(X(X-1)) + \mu - \mu^2$. We can find the first term on the RHS by using the second derivative of $G_X(t)$:

$$G_X''(t) = 16e^{4(t-1)} \Rightarrow \quad \text{E}(X(X-1)) = G_X''(1) = 16$$

$$\Rightarrow \quad \sigma^2 = \text{Var}(X) = 16 + \mu - \mu^2 = 16 + 4 - 4^2 = 4$$ ◄

As the example shows, probability generating functions can be very useful (but not always!) in the derivation of expected values and variances. In addition, they can also be very helpful in the determination of the distribution of a function, say *Y,* of one or more random variables. That is because a probability generating function is *unique* for a specific probability distribution, so two different distributions never have the same probability generating function! If we are able, in any way whatsoever, to determine the probability generating function of the random variable *Y*, and if we recognize the resulting function as the probability generating function of one of the common probability distributions (discussed in section 3.7), then we have actually identified the probability distribution of *Y* !

3.6 Sums of random variables

Convolution formula

Suppose *X* and *Y* are two independent random variables, e.g. *X* is the number of dots when throwing an unbiased die, and *Y* is the result when throwing the die again. Just as in this example, we will assume that *X* and *Y* can only have nonnegative integers as possible values. An interesting question is how to find a general formula for the probability that $P(X + Y = k)$ (for $k = 0, 1, 2, \ldots$). We can see that the outcome $X + Y = k$ can only occur when $X = i$ and $Y = k - i$ for one or more values of *i*, with $i = 0, 1, 2, \ldots, k$. That leads to:

$$P(X + Y = k) = \sum_{i=0}^{k} P(X = i \text{ and } Y = k - i)$$

But since the two events $X = i$ and $Y = k - i$ come from two independent trials, which implies that the two events are independent, we know (see section 2.5.2):

$$P(X = i \text{ and } Y = k - i) = P(X = i)\, P(Y = k - i)$$

This results in the so-called convolution formula:

$$P(X + Y = k) = \sum_{i=0}^{k} P(X = i)\, P(Y = k - i)$$

Example 3.12

What is the probability of throwing a total of 8 dots with two unbiased dice? Define *X* as the number of dots on die 1, and *Y* as the number on die 2.

$$P(X + Y = 8) = \sum_{i=0}^{8} P(X = i)\, P(Y = 8 - i)$$

$$= \sum_{i=2}^{6} P(X = i)\, P(Y = 8 - i)$$ (check carefully why the adjusted summation indexes are correct!!)

$$= 5 \cdot \frac{1}{6} \cdot \frac{1}{6} = \frac{5}{36}$$ ◄

In order to study the distribution of the sum $X + Y$ of two random variables in more general situations, we actually need more advanced theory, which will be discussed in the next course, Probability Theory and Statistics 2. However, because some results to be discussed there are very helpful in this course already, we will give some basic results here (without proofs). Also note that Theorem 3.10 and Theorem 3.11 mention the concept of *independent* random variables. A proper definition of independency of random variables will have to wait until the next course as well; for the time being, it is sufficient to think of independent random variables as random variables which are defined on

different trials of the same experiment, or even on different experiments. On the other hand, when for example the experiment is to select at random an adult male from the Dutch population, and if we define X as the weight and Y as the height of the selected person, then X and Y will *not* be independent.

Theorem 3.9 *(B&E, Th. 5.2.2)*

If X and Y are random variables, then

$$E(X + Y) = E(X) + E(Y)$$

Theorem 3.10 *(≈B&E, Th. 5.2.6)*

If X and Y are independent random variables, then

$$Var(X + Y) = Var(X) + Var(Y)$$

Theorem 3.11

If X and Y are independent random variables, then

$$G_{X+Y}(t) = G_X(t) \cdot G_Y(t).$$

In other words: the probability generating function of $X + Y$ is equal to the product of the two probability generating functions of X and Y.

Note the important difference: the result in the first theorem is **always** valid, the other two only when the random variables are independent.

3.7 Special discrete distributions

(B&E, Page 91-108)

3.7.1 The discrete uniform distribution

A discrete uniform distribution is characterised by a finite number of possible outcomes, all equally likely to be observed (hence the name 'uniform'). That probability is therefore equal to 1 divided by the number of possible outcomes. In addition to this, most authors limit the set of possible outcomes to a number of successive integers, for example 5, 6, 7, 8 and 9, each with a probability of 1/5. And some books, like Bain & Engelhardt, define this distribution only for numbers starting at 1. Here we do it slightly more general:

Definition 3.7 *(B&E, Page 107)*

A discrete random variable X has a **discrete uniform distribution** (or discrete homogeneous distribution; Dutch: discreet uniforme verdeling) on the set $\{a, a+1, a+2, \ldots, b\}$ (notation: $X \sim$ DU(a, b)) if its probability distribution function is given by:

$$f(x) = \frac{1}{b-a+1} \quad \text{for } x = a, a+1, a+2, ..., b$$

where a and b are integers with $a \leq b$.

Example 3.13

Experiment: Throw an unbiased die. Define X as the number of dots. Then $X \sim$ DU(1, 6), with $f(x) = 1/6$ for $x = 1, 2, 3, 4, 5, 6$. ◄

Expected value, variance and probability generating function

We will determine here the expected value, variance and probability generating function of X when

$X \sim$ DU(1, N), so with $f(x) = 1/N$ for x = 1, 2, …, N. We will often need knowledge of mathematical series to find those results, see Appendix A.3. In the exercises, you will be asked to generalise these results to DU(a, b)-distributed random variables (see Problem 3.30)

$$\mathrm{E}(X) = \sum_{i=1}^{N} \frac{i}{N} = \frac{1}{N}(1 + 2 + ... + N) = \frac{1}{N}\frac{1}{2}N(N+1) = \frac{N+1}{2}$$

$$\mathrm{E}(X^2) = \sum_{i=1}^{N} \frac{i^2}{N} = \frac{1}{N}(1 + 4 + ... + N^2)$$

$$= \frac{1}{N}\frac{1}{6}N(N+1)(2N+1) = \frac{(N+1)(2N+1)}{6}$$

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - (\mathrm{E}(X))^2$$

$$= \frac{(N+1)(2N+1)}{6} - \frac{(N+1)(N+1)}{4} = \frac{(N^2-1)}{12}$$

$$G_X(t) = \sum_{i=1}^{N} t^i \frac{1}{N} = \frac{1}{N}(\sum_{i=0}^{N} t^i - 1) = \frac{1}{N}(\frac{t^{N+1}-1}{t-1} - \frac{t-1}{t-1}) = \frac{t(t^N-1)}{N(t-1)}$$

3.7.2 The Bernoulli distribution

Often, each trial of an experiment can result in only two possible outcomes, one of which is usually called a 'success', while the other is called a 'failure'. Such a trial is then called a **Bernoulli trial**. It is a common procedure to define a random variable X for this very simple sample space by assigning the value 0 when a trial results in a failure and the value 1 when a trial results in a success. The probability of success is denoted by p and the probability of failure is often denoted by q (or simply by $1-p$).

Definition 3.8 *(B&E, Page 91)*

A discrete random variable X has a **Bernoulli distribution** with parameter p ($0 \le p \le 1$) if its probability distribution function is given by:

$$f(x) = p^x q^{1-x} \quad \text{for } x = 0,\ 1$$

where $0 \le p \le 1$, and $q = 1 - p$.

Note that the formula above simply leads to $f(0) = q$ and $f(1) = p$.

Expected value, variance and probability generating function

Notice that $\mathrm{E}(X) = 0 \cdot f(0) + 1 \cdot f(1) = 0 \cdot q + 1 \cdot p = p$

And $\mathrm{E}(X^2) = 0^2 \cdot f(0) + 1^2 \cdot f(1) = p$

The variance is therefore:

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - (\mathrm{E}(X))^2 = p - p^2 = p(1-p) = pq$$

The probability generating function is also very simple:

$$G_X(t) = \mathrm{E}(t^X) = \sum_{k=0}^{1} t^k \ P(X = k) = q + pt$$

(Check for yourself how you can use this function to determine the expected value and the variance in an alternative way, see section 3.5).

3.7.3 The binomial distribution

Again we think of an experiment with only two possible outcomes, success and failure. When we perform such an experiment n times, we can count the number of successes in those n trials. Whenever

the probability of success remains the same for each trial, irrespective of the outcomes of the previous trials (i.e. we are dealing with independent Bernoulli trials), then the distribution of the number of successes is called a binomial distribution. When we think of an experiment as drawing at random an element from a population, then we can see that we are dealing with independent Bernoulli trials when either: 1. the population size is finite and each selected element is replaced before the next trial (because at the next draw, the population is then exactly equal to the population at the very beginning), or 2. the population size in infinite (because in that case drawing one element from the population does not change in any way the proportion of successes in the population). Below is an example of the latter case:

Example 3.14

Assume a certain student simply guesses each of the answers in a test with ten multiple-choice questions. At each guess, the student will guess the correct answer (S=Success) with probability 0.25, and he will guess the incorrect answer (F=Failure) with probability 0.75. What is the probability of guessing exactly two correct answers out of the total of ten? First, we will determine the probability of answering *the first two* questions correct, followed by eight incorrect answers (SSFFFFFFFF). That probability is (see Definition 2.2):

$$0.25 \cdot 0.25 \cdot 0.75 \cdot 0.75 \cdot 0.75 \cdot 0.75 \cdot 0.75 \cdot 0.75 \cdot 0.75 \cdot 0.75 = 0.25^2 \cdot 0.75^8$$

But this same probability applies to each sequence of two correct and eight incorrect answers (e.g. SFFSFFFFFF or FFFFSFFFSF). In section 2.7.3 , we have seen that the number of sequences is equal to $\binom{10}{2}$. This results in the total probability of $\binom{10}{2} 0.25^2 \cdot 0.75^8 = 0.282$ ◄

Definition 3.9 *(B&E, Page 92)*

A discrete random variable X has a **binomial distribution** with parameters n and p (notation: $X \sim \text{BIN}(n, p)$) if its probability distribution function is given by:

$$f(x) = P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} = \binom{n}{x} p^x q^{n-x} \qquad x = 0, 1, \ldots, n$$

where $0 \le p \le 1$, $q = 1 - p$, and n an integer ≥ 1.

Note: The binomial coefficient $\binom{n}{x}$ gives the number of possible sequences of x successes and $n - x$ failures, each with the probability $p^x q^{n-x}$.

Note also that the Bernoulli distribution is identical to a binomial distribution with parameter $n = 1$. Most tables of binomial distributions only show the cumulative probabilities:

$$P(X \le k) = \sum_{j=0}^{k} \binom{n}{j} p^j q^{n-j} \qquad \text{for } k = 0, 1, ..., n \text{ [or } n - 1]$$

For example for ***n* = 10**:

							p								
k	**.01**	**.05**	**.10**	**.20**	**.25**	**.30**	**.40**	**.50**	**.60**	**.70**	**.75**	**.80**	**.90**	**.95**	**.99**
0	.904	.599	.349	.107	.056	.028	.006	.001	.000	.000	.000	.000	.000	.000	.000
1	.996	.914	.736	.376	.244	.149	.046	.011	.002	.000	.000	.000	.000	.000	.000
2	1.000	.988	.930	.678	.526	.383	.167	.055	.012	.002	.000	.000	.000	.000	.000
3	1.000	.999	.987	.879	.776	.650	.382	.172	.055	.011	.004	.001	.000	.000	.000
4	1.000	1.000	.998	.967	.922	.850	.633	.377	.166	.047	.020	.006	.000	.000	.000
5	1.000	1.000	1.000	.994	.980	.953	.834	.623	.367	.150	.078	.033	.002	.000	.000
6	1.000	1.000	1.000	.999	.996	.989	.945	.828	.618	.350	.224	.121	.013	.001	.000
7	1.000	1.000	1.000	1.000	1.000	.998	.988	.945	.833	.617	.474	.322	.070	.012	.000
8	1.000	1.000	1.000	1.000	1.000	1.000	.998	.989	.954	.851	.756	.624	.264	.086	.004
9	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.999	.994	.972	.944	.893	.651	.401	.096

(Why can the line $k = 10$ be omitted from this table without any problem?)
Individual probabilities can be found by taking the difference between two successive cumulative probabilities:

$$P(X = k) = P(X \le k) - P(X \le k-1)$$

The following relation can also be very helpful: $P(X \ge k) = 1 - P(X \le k-1)$

Example 3.15

See Example 3.14. Use the table above to determine the probability of exactly two correct answers.

$$P(X = 2) = P(X \le 2) - P(X \le 1) = 0.526 - 0.244 = 0.282$$

And the probability of at least two correct answers:

$$P(X \ge 2) = 1 - P(X < 2) = 1 - P(X \le 1) = 1 - 0.244 = 0.756$$ ◄

Probability generating function

Using the definition for the probability generating function, we find for $X \sim \text{BIN}(n,p)$:

$$G_X(t) = \text{E}(t^X) = \sum_{k=0}^{n} t^k \; P(X = k) = \sum_{k=0}^{n} t^k \binom{n}{k} p^k q^{n-k}$$

$$= \sum_{k=0}^{n} \binom{n}{k} (pt)^k q^{n-k} = (pt + q)^n$$

Note that the last equality above follows directly from the Binomial Theorem (Theorem 2.12).

Expected Value

By intuition, most people will agree that the expected number of successes in n trials is equal to np. Let us check if that is also the result by applying the definition of the expected value.

$$\text{E}(X) = \sum_{k=0}^{n} k \; P(X = k) = \sum_{k=0}^{n} k \binom{n}{k} p^k q^{n-k} = \sum_{k=1}^{n} k \, \frac{n!}{k!(n-k)!} p^k q^{n-k} =$$

(because $k = 0$ results in a term equal to 0. For the next step, we can cancel out a factor k from both numerator and denominator, giving:)

$$= \sum_{k=1}^{n} \frac{n(n-1)!}{(k-1)!(n-k)!} p^k q^{n-k} = np \sum_{k=1}^{n} \frac{(n-1)!}{(k-1)!(n-k)!} p^{k-1} q^{n-k} = np \sum_{k=1}^{n} \binom{n-1}{k-1} p^{k-1} q^{n-k}$$

(A factor np was written in front of the summation sign. In the next step, we substitute k by $h + 1$.)

$$= np \sum_{h=0}^{n-1} \binom{n-1}{h} p^h q^{n-1-h} = np$$

(Note that the last sum is equal to 1, because it is just the sum of probabilities of a binomial distribution with $n - 1$ trials and same success probability p.)

But in this case, it would have been a lot simpler to use the probability generating function:

$$G_X'(t) = \frac{d}{dx}(pt + q)^n = np(pt + q)^{n-1} \quad \Rightarrow \quad \text{E}(X) = G_X'(1) = np(p + q)^{n-1} = np$$

Variance

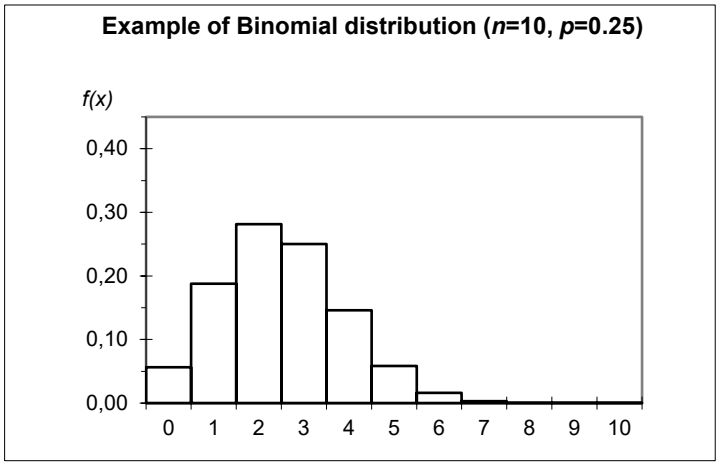
We will again use the probability generating function:

$$G_X''(t) = n(n-1)p^2(pt + q)^{n-2} \quad \Rightarrow \quad \text{E}(X(X-1)) = G_X''(1) = n(n-1)p^2$$

$$\text{Var}(X) = \text{E}(X^2) - (\text{E}(X))^2 = \text{E}(X(X-1)) + \text{E}(X) - (\text{E}(X))^2 =$$

$$= n(n-1)p^2 + np - (np)^2 = np(1-p) = npq$$

Therefore, the standard deviation is $\sigma = \sqrt{npq}$.



3.7.4 The hypergeometric distribution

Again, we will assume that all elements within a population represent either a success or a failure and we are interested in the number of successes X in a random sample of n elements. However, now we will consider a finite population (consisting of $N < \infty$ elements) where the elements will *not* be replaced into the population after each draw (drawing without replacement). If we denote the number of successes in the population by M, then of course the number of failures is $N - M$. At the first draw, the probability of success is clearly M/N. However, the probability of success at the second trial now depends on the outcome at the first trial: if the first trial resulted in success, then only M–1 of the remaining N–1 elements are successes, so the probability at the second draw becomes (M–1)/(N–1), while this probability will be $M/(N–1)$ if the first trial was a failure. In other words, the binomial model does not apply, since the trials are now no longer independent.

Example 3.16

A group of 70 students consisting of 40 male and 30 female students will be randomly divided into two tutorial groups of 35 students each. What is the probability that the first tutorial group will count exactly 20 male students? We can find this probability using the classical definition of probability (see 2.4.1). The denominator of this probability is the total number of ways to select 35 students from a group of 70, which is $\binom{70}{35}$. In the numerator, we multiply (because of the multiplication principle) the number of ways to select 20 male students out of a total of 40 male students by the number of ways to select the remaining 35 – 20=15 students out of a total of 30 female students, so $\binom{40}{20}\binom{30}{15}$. The requested probability is $\binom{40}{20}\binom{30}{15}\Big/\binom{70}{35} = 0.191$. ◄

Definition 3.10 *(B&E, Page 95)*

A discrete random variable X has a **hypergeometric distribution** with parameters n, M and N (notation: $X \sim \text{HYP}(n, M, N)$) if its probability distribution function is given by:

$$f(x) = P(X = x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} = \frac{\binom{n}{x}\binom{N-n}{M-x}}{\binom{N}{M}} \quad \text{for } x = 0,\ 1,\ 2,\ ...,\ n$$

where n, M and N are positive integers with $n \leq N$ and $M \leq N$.

By simply replacing each binomial coefficient by an expression consisting of factorials, we can easily check the equality of both expressions above (see also exercise 3.54). In other words, the roles of n and M are interchangeable.

Also notice that the support for X is often only a subset of the set $\{0, 1, \ldots, n\}$. For example, it is not possible that the number of successes x in the sample will ever be greater than the number of successes M in the population. However, the above formula is still valid, because the binomial coefficient $\binom{M}{x}$ is by matter of definition equal to 0 whenever $x > M$ (see section 2.7.3).

Hypergeometric probabilities are seldom presented in the form of tables, simply because of the fact that there are no less than three parameters, so we would need a very thick table book for just the hypergeometric distribution alone!

Relation with the binomial distribution

If we would consider taking a random sample *with* replacement, then the probability of success at each trial would be equal to $p=M/N$. In that case, the probability of x successes can be found by the binomial pdf:

$$P(X=x)=\binom{n}{x}\left(\frac{M}{N}\right)^x\left(\frac{N-M}{N}\right)^{n-x}=\binom{n}{x}p^x(1-p)^{n-x}$$

Of course, these probabilities are different from the hypergeometric probabilities. But when the size of the sample n is very small compared to the total population size N, then we would expect that it hardly matters whether we sample with or without replacement. When we take a very small random sample without replacement from a very large population, then the probability of success at any draw will be a number very close to the original value of M/N. Indeed, we can show (see exercise 3.57) that a hypergeometric probability will become almost equal to a binomial probability whenever $N >> n$. Because of the fact that the roles of n and M can be interchanged, the same is true whenever $N >> M$. This means that we may use the binomial distribution as an approximation for the hypergeometric distribution when either n or M is much smaller than N. (How much smaller? Some books use the rule of thumb that such an approximation is reasonable when n or $M < N/20$).

The relationship between the two distributions also shows up when we determine the expected value and the variance. Because, unfortunately, it is not possible to obtain a usable expression for the probability generating function for the hypergeometric distribution, we can only determine those values using the pdf, as seen below.

Expected value

It seems very reasonable to 'expect' that the mean of a random variable with a hypergeometric distribution is equal to $n \times (M/N)$ (as compared to the mean np for the binomial distribution). We will check this here:

$$\mathrm{E}(X)=\sum_{k=0}^{n}k\,P(X=k)=\sum_{k=0}^{n}k\,\frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}=\sum_{k=1}^{n}k\,\frac{\frac{M!}{k!(M-k)!}\binom{N-M}{n-k}}{\binom{N}{n}}=$$

because $k = 0$ results in a first term equal to 0. Now the factor k in numerator and denominator cancel each other out, and we bring a factor nM/N in front of the summation sign:

$$=\sum_{k=1}^{n}\frac{\frac{M!}{(k-1)!(M-k)!}\binom{N-M}{n-k}}{\frac{N!}{n!(N-n)!}}=\frac{nM}{N}\sum_{k=1}^{n}\frac{\frac{(M-1)!}{(k-1)!(M-k)!}\binom{N-M}{n-k}}{\frac{(N-1)!}{(n-1)!(N-n)!}}=\frac{nM}{N}\sum_{k=1}^{n}\frac{\binom{M-1}{k-1}\binom{N-M}{n-k}}{\binom{N-1}{n-1}}=$$

In the next step, we substitute k by $h + 1$:

$$= \frac{nM}{N} \sum_{h=0}^{n-1} \frac{\binom{M-1}{h}\binom{N-1-(M-1)}{n-1-h}}{\binom{N-1}{n-1}} = \frac{nM}{N} = n\frac{M}{N}$$

The last sum is taken over all probabilities from a hypergeometric distribution with $n-1$ draws from a population of $N-1$ elements with $M-1$ successes, and we know that such a sum (a sum of all probabilities for a discrete random variable) must be equal to 1.

An alternative approach to determine the expected value could be as follows: we define a sequence of random variables $Y_1, Y_2, ..., Y_n$ with $Y_i = 1$ if the i-th trial results in success and 0 otherwise. Then $X = Y_1 + Y_2 + ... + Y_n$ which gives $\mathrm{E}(X) = \sum_{i=1}^{n} \mathrm{E}(Y_i)$. It is obvious that $\mathrm{E}(Y_1) = 1 \cdot \frac{M}{N} + 0 \cdot \frac{(N-M)}{N} = \frac{M}{N}$.

To find $\mathrm{E}(Y_2)$, we can first determine $P(Y_2 = 1)$ by considering what could have happened in the first draw:

$$P(Y_2 = 1) = P(Y_2 = 1 \mid Y_1 = 0) \cdot P(Y_1 = 0) + P(Y_2 = 1 \mid Y_1 = 1) \cdot P(Y_1 = 1)$$
$$= \frac{M}{N-1} \cdot \frac{N-M}{N} + \frac{M-1}{N-1} \cdot \frac{M}{N} = \frac{M}{N}$$

Thus, $\mathrm{E}(Y_2) = M/N$, and by similar argument: $\mathrm{E}(Y_i) = \frac{M}{N}$ for i = 1, 2, …. .

So, even though the trials are not independent, the expected values are still the same for each trial (but of course only as long as the outcomes of the first trials are unknown; this can also be seen by the following thought-experiment: a mechanised robot with n arms ($n \le N$) draws n-elements from a population at the same time. There is really no reason why the probability of success would be different from arm to arm.)

Finally, this gives us the result $\mathrm{E}(X) = \sum_{i=1}^{n} \mathrm{E}(Y_i) = n\mathrm{E}(Y_1) = n\frac{M}{N}$.

Variance

We can use a similar approach as above to find the variance of X, i.e. by writing $X = Y_1 + Y_2 + ... + Y_n$. Since the trials are not independent, we cannot use Theorem 3.10; we need more advanced theory (on multivariate random variables and covariances) which will only be discussed in the next course. Alternatively, we could first determine $\mathrm{E}(X(X-1))$ by evaluating this as a sum using the hypergeometric pdf. Next, the variance follows by $\mathrm{Var}(X) = \mathrm{E}(X^2) - (\mathrm{E}X)^2 = \mathrm{E}(X(X-1)) + \mathrm{E}(X) - (\mathrm{E}X)^2$. This is a lot of work, and in itself not so very interesting. However, the result is interesting indeed:

$$\mathrm{Var}(X) = \frac{nM}{N}\frac{(N-M)}{N}\frac{(N-n)}{(N-1)}$$

When we replace M/N by p and $(N-M)/N$ by q, then the variance can be written as: $npq\frac{N-n}{N-1}$.

We recognise the factor npq, which is the variance for the binomial distribution. The factor $\frac{N-n}{N-1}$ is called the 'finite population correction factor'. For any value of $n > 1$, this factor will be less than 1, which actually means that the variance is less in random samples without replacement as compared to samples with replacement. The more elements of the population are sampled without replacement, the smaller the variance. Finally, when the sample is as large as the population size (so when $n = N$), the variance of the number of successes will even be 0! This can be explained very simply, because in the latter case we have sampled the complete population, so the value of X will be equal to M with absolute certainty.

3.7.5 The geometric distribution

Again we will consider a sequence of independent Bernoulli trials. Suppose however that we are now not interested in the number of successes during a fixed number of trials (like we did when discussing the binomial distribution), but instead in the number of trials needed to observe the first success.

Example 3.17

Experiment: throw an unbiased die until the first time that 6 dots appear. Define X as the total number of throws (so including the last throw which resulted in 6 dots). What is the probability that the first '6' will be thrown at the fourth throw, or $P(X = 4)$? This can happen only when the first three throws all resulted in outcomes not equal to '6', followed at the fourth throw by '6'. The probability of $X = 4$ is therefore 5/6*5/6*5/6*1/6 = $1/6 \cdot (5/6)^3 = 0.0965$ ◄

Definition 3.11 *(B&E, Page 99)*

A discrete random variable X has a **geometric distribution** with parameter p (notation: $X \sim \text{GEO}(p)$) if its probability distribution function is given by:

$$f(x) = P(X = x) = pq^{x-1} \qquad x = 1, 2, \ldots$$

where $0 \le p \le 1$, and $q = 1 - p$.

It is very simple to check that the sum over all probabilities is 1 (see also Appendix A.3):

$$\sum_{x=1}^{\infty} pq^{x-1} = p(1 + q + q^2 + \cdots) = p\left(\frac{1}{1-q}\right) = \frac{p}{p} = 1$$

The CDF for this distribution does not need to be written as a sum, but is very simple to find:

$$F(x) = P(X \le x) = 1 - P(X > x) = 1 - q^x$$

(because $X > x$ only happens when the first x trials passed without any success)

Since both pdf and CDF are so easy to evaluate, most books do not show any tables for this distribution. Please be warned that some authors and some computer packages define the geometric distribution in a slightly different way. Instead of counting the number of trials needed to observe the first success, they count the number of trials *before* the first success occurs. This results in $f(x) = pq^x$ for $x = 0, 1, \ldots$. As long as one is aware of this difference, it should not cause any problems.

Probability generating function

The probability generating function for the geometric distribution is

$$G_X(t) = \text{E}(t^X) = \sum_{k=1}^{\infty} t^k \; P(X = k) \; = \; \sum_{k=1}^{\infty} t^k \; pq^{k-1}$$

$$= \; pt \sum_{k=1}^{\infty} (qt)^{k-1} \; = \frac{pt}{1 - qt}$$

Notice that in the last line above, the sum of terms on the left-hand side only converges when $|qt| < 1$, so when $-1/q < t < 1/q$ (see Appendix A.3). So, strictly speaking, the result $pt/(1-qt)$ is only correct when this condition is satisfied, and, one might argue, that should be clearly indicated. However, since we are only ever interested in (the behaviour of) the probability generating function for values of t around 0 and around 1, this condition poses no restrictions at all to the use of the result $pt/(1-qt)$.

Expected value

The expectation can be determined as follows:

$$\begin{aligned}\mathrm{E}(X) &= \sum_{k=1}^{\infty} k\ P(X=k) \\ &= \sum_{k=1}^{\infty} k\ pq^{k-1} = \frac{p}{q}\sum_{k=1}^{\infty} kq^k \\ &= \frac{p}{q}\frac{q}{(1-q)^2} = \frac{p}{p^2} = \frac{1}{p}\end{aligned}$$

Again, we used a well-known series here (see Appendix A.3).

Or using the probability generating function:

$$G_X'(t) = \frac{d}{dt}\left(\frac{pt}{1-qt}\right) = \frac{p(1-qt)-pt(-q)}{(1-qt)^2} = \frac{p}{(1-qt)^2}$$

So $\mathrm{E}(X) = G_X'(1) = \dfrac{p}{(1-q)^2} = \dfrac{1}{p}$

This is indeed according to the intuition of most people; for example, when the probability of '6 dots' is 1/6, then the average number of attempts needed to throw '6 dots' will be equal to 6.

Variance

We will only here show the use of the probability generating function to determine the variance:

$$G_X''(t) = \frac{d}{dt}\left(\frac{p}{(1-qt)^2}\right) = \frac{2pq}{(1-qt)^3}$$

$$\mathrm{E}(X(X-1)) = G_X''(1) = \frac{2pq}{(1-q)^3} = \frac{2q}{p^2}$$

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - (\mathrm{E}X)^2 = \mathrm{E}(X(X-1)) + \mathrm{E}X - (\mathrm{E}X)^2 = \frac{2q}{p^2} + \frac{1}{p} - \frac{1}{p^2} = \frac{1-p}{p^2} = \frac{q}{p^2}$$

Memoryless property

The geometric distribution is the only discrete probability distribution with the memoryless property, which can be written as $P(X > k+j \mid X > k) = P(X > j)$ (with k and j positive integers). So, if we know that the first k trials all resulted in failures ($X > k$), then the conditional probability of not obtaining any success in *next* j trials is equal to the **un**conditional probability of not obtaining any success in j trials. For example, knowing that you have thrown a die already k times without observing '6 dots' does not affect in any way the distribution of the number of additional throws needed until the first time '6 dots' appear. In other words, no matter how many failures have already occurred, the next success does not get any closer (nor further removed).

$$\begin{aligned}P(X > k+j \mid X > k) &= \frac{P(X > k+j \cap X > k)}{P(X > k)} \\ &= \frac{P(X > k+j)}{P(X > k)} = \frac{q^{k+j}}{q^k} = q^j = P(X > j)\end{aligned}$$

3.7.6 The negative binomial distribution

The geometric distribution dealt with the number of independent Bernoulli trials needed until the first success. This can easily be extended to the number of trails needed until the r-th success.

Example 3.18

Experiment: throw a unbiased die until the *second* time 6 dots appear. Define X as the total number of throws needed. What is the probability that the second '6' will be thrown at the eighth throw, or $P(X = 8)$? This can happen only when the first seven throws resulted in exactly one '6', followed at the eighth throw by '6'. Note that the probability of exactly one success in seven trials is just a probability from the binomial distribution. Therefore, the probability of $X = 8$ is:

$$\binom{7}{1}\left(\frac{1}{6}\right)^1\left(\frac{5}{6}\right)^6 \cdot \frac{1}{6} = \binom{7}{1}\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right)^6 = 0.065$$ ◄

Definition 3.12 *(B&E, Page 101)*

A discrete random variable X has a **negative binomial distribution** with parameters r and p (notation: $X \sim \text{NB}(r, p)$ or $X \sim \text{NEGBIN}(r, p)$), if its probability distribution function is given by:

$$f(x) = P(X = x) = \binom{x-1}{r-1} p^r q^{x-r} \qquad x = r, r+1, \ldots$$

where $0 \le p \le 1$, $q = 1 - p$ and r integer > 0.

Explanation: The binomial coefficient in the definition above is equal to the number of possible sequences of $r - 1$ successes in the first $x - 1$ experiments. Furthermore, there should be a total of r successes and $x - r$ failures, and therefore these numbers are to be found in the powers of the probabilities p and q.

Remark 1. Of course, the sum of all probabilities must be equal to 1. To verify this, we need knowledge about so-called binomial series; we will not do that here (see possibly B & E, p. 102).

Remark 2. Some statistics textbooks (and EXCEL!) count the number of *failures* prior to the r-th success in the definition of the negative binomial variable. Be aware of this difference in definition!

Note that the geometric distribution is a special case of the negative binomial distribution (with $r = 1$). A negative binomial random variable X can also be viewed as the sum of r independent geometrically distributed random variables, so $X = Y_1 + Y_2 + \ldots + Y_r$, each with the same parameter p (Y_1 is the number of trials needed until the first success, Y_2 the number of trials needed until the second success when we start counting after the first success, etc.)

Probability generating function

$$G_X(t) = \text{E}(t^X) = \sum_{k=r}^{\infty} t^k \, P(X = k) = \sum_{k=r}^{\infty} t^k \binom{k-1}{r-1} p^r q^{k-r}$$

$$= \sum_{k=r}^{\infty} \binom{k-1}{r-1} (pt)^r (qt)^{k-r} =$$

Comparing the last expression above with the pdf for the negative binomial distribution, we can notice that it starts to look again like the sum over all probabilities for the negative binomial distribution with probability of failure qt (provided that $0 \le t < 1/q$). Only the factor with pt is not correct. However, we can rewrite the above expression as:

$$= \frac{(pt)^r}{(1-qt)^r} \sum_{k=r}^{\infty} \binom{k-1}{r-1} (1-qt)^r (qt)^{k-r} = \frac{(pt)^r}{(1-qt)^r}$$

The last equality follows because we know that the sum of all probabilities for any negative binomial distribution (here with parameters r and $1–qt$) must be equal to 1.

A more simple way of finding this probability generating function makes use of $X = Y_1 + Y_2 + ... + Y_r$ with $Y_i \sim \text{GEO}(p)$ and independent from each other (see above). By applying Theorem 3.11. repeatedly, we get:

$$G_X(t) = G_{Y_1}(t)G_{Y_2}(t)\cdots G_{Y_r}(t) = \frac{pt}{1-qt}\cdot\frac{pt}{1-qt}\cdot\ \ \cdots = \left(\frac{pt}{1-qt}\right)^r .$$

Expected value and variance

In the previous section, we found the expectation of the number of trials needed until the first success to be $1/p$. It seems obvious that the expectation of the number of trials needed until the r-th success should be r times $1/p = r / p$. This is indeed what follows directly from Theorem 3.9 and the fact that X can be written as the sum of independent geometric random variables, $X = Y_1 + Y_2 + ... + Y_r$.

In a very similar way, it can be shown with the help of Theorem 3.10 and the variance for the geometric distribution that $\text{Var}(X) = rq / p^2$.

Otherwise, of course, we could have used the probability generating function as well (try it).

Relationship between the binomial and the negative binomial distribution

Cumulative probabilities for the negative binomial distribution can be derived directly from cumulative probabilities for the binomial distribution. Whenever the number of trials needed to obtain the r-th success is less than or equal to n , this must logically mean that the number of successes Y in the first n trials must be at least r. Since $Y \sim \text{BIN}(n , p)$, we can write the relationship as follows:

$$P(X \le n \mid X \sim \text{NEGBIN}(r, p)\ ) = P(Y \ge r \mid Y \sim \text{BIN}(n , p)\ )$$

Or, equivalently:

$$P(X > n \mid X \sim \text{NEGBIN}(r, p)\ ) = P(Y < r \mid Y \sim \text{BIN}(n , p)\ )$$

Remark: When we use a notation like $P(Y \ge r \mid Y \sim \text{BIN}(n , p)\ )$, we simply mean $P(Y \ge r)$ given that $Y \sim \text{BIN}(n , p)$. This notation works intuitively fine, and is very convenient when we are dealing with several different probability distributions. However, note that we cannot compute such a probability by the rule $P(A \mid B) = \dfrac{P(A \cap B)}{P(B)}$, because "$Y \sim \text{BIN}(n , p)$" is not an event (= a subset of the sample space).

Example 3.19

Suppose the probability of success p is 0.3. What is the probability that a maximum of 20 trials are needed to obtain the fifth success? We define X as the number of trials required until the fifth success, so $X \sim \text{NEGBIN}(5, 0.3)$ and we need to determine $P(X \le 20)$. This probability will be equal to the (binomial) probability of at least 5 successes in 20 trials, so $P(Y \ge 5)$ with $Y \sim \text{BIN}(20, 0.3)$.

$$P(X \le 20) = P(Y \ge 5) = 1 - P(Y \le 4)$$

This last probability can be found in a standard binomial table, so $P(X \le 20) = 1 - 0.2375 = 0.7625$.

And if we need to find $P(X = 20)$, we can write :

$$P(X = 20) = P(X \le 20) - P(X \le 19)$$
$$= P(Y \ge 5 \mid Y \sim \text{BIN}(20, 0.3)) - P(Y \ge 5 \mid Y \sim \text{BIN}(19, 0.3))$$
$$= 0.7625 - 0.7178 = 0.0447.$$

◄

3.7.7 The Poisson distribution

(B&E, Page 103-107)

The Poisson distribution often shows up in situations where we are dealing with the number of events occurring in a fixed time or space interval, if these events occur with a certain average rate and independently of the time (or space) since the last event occurred. Examples:

- the number of traffic accidents along a specific road section per year;
- the number of lightning strikes in Amsterdam in August;
- the number of people deaths from tetanus in the Netherlands per year;
- the number of flaws per square metre of clothing.

The only parameter of this distribution is the *expected number of times* that the event will occur in the chosen interval, notated by μ. In contrast with the previous distributions, we will not start by deriving the pdf; Later, in Theorem 3.12, we will show that this pdf can be seen at the limiting case of a binomial distribution.

Definition 3.13

A discrete random variable X has a **Poisson distribution** with parameter μ (notation: $X \sim \text{POI}(\mu)$) if its probability distribution function is given by:

$$f(x) = P(X = x) = e^{-\mu}\,\frac{\mu^x}{x!} \quad \text{for } x = 0,\ 1,\ 2,\ ....$$

where $\mu > 0$.

Example 3.20

If the number of deaths X from tetanus is on average 1.5 per year, then the probability of 0 and 3 deaths respectively in a specific year is (assuming X is Poisson distributed).

$$P(X = 0) = e^{-1.5}\,\frac{1.5^0}{0!} = e^{-1.5} = 0.223130$$

$$P(X = 3) = e^{-1.5}\,\frac{1.5^3}{3!} = 0.22313016 \times \frac{3.375}{6} = 0.125511$$ ◄

Most tables for the Poisson distribution show cumulative probabilities for different values of μ, so

$$F_X(x) = P(X \le x) = \sum_{j=0}^{x} e^{-\mu}\,\frac{\mu^j}{j!} \qquad (x = 0, 1, 2, \ldots)$$

Although the support for the Poisson distribution is the set of natural numbers, the probability of observing an outcome much larger than the value of μ quickly becomes almost negligible, which restricts the number of rows shown in most Poisson tables. Also, most tables do not go beyond μ = 15, because we will see in the next chapter that we can approach the Poisson probabilities for larger values of μ very well using the normal distribution. Note that individual probabilities can be calculated using:

$$P(X = k) = P(X \le k) - P(X \le k - 1).$$

Probability generating function

We derive:

$$G_X(t) = \text{E}(t^X) = \sum_{k=0}^{\infty} t^k\ \text{P}(X = k) = \sum_{k=0}^{\infty} t^k\ e^{-\mu}\,\frac{\mu^k}{k!} = \sum_{k=0}^{\infty} e^{-\mu}\,\frac{(t\mu)^k}{k!} = e^{-\mu}\sum_{k=0}^{\infty}\frac{(t\mu)^k}{k!} =$$

$$= e^{-\mu+t\mu} = e^{\mu(t-1)} \qquad (\text{Using } e^y = \sum_{k=0}^{\infty}\frac{y^k}{k!}, \text{ see Appendix A.3 })$$

Expected value and variance

Note that we have used the symbol μ for the parameter of the Poisson distribution, the same symbol we introduced for the population mean. So let us check if $\mathrm{E}(X)$ is indeed equal to this parameter μ.

$$\mathrm{E}(X) = \sum_{k=0}^{\infty} k\, P(X=k) = \sum_{k=0}^{\infty} k\, e^{-\mu}\frac{\mu^k}{k!} = e^{-\mu}\sum_{k=1}^{\infty}\frac{\mu^k}{(k-1)!}$$

$$= \mu e^{-\mu}\sum_{k=1}^{\infty}\frac{\mu^{k-1}}{(k-1)!} = \mu e^{-\mu}e^{\mu} = \mu$$

We can use the probability generating function as well to find the expectation and the variance. The first two derivatives are:

$$G_X'(t) = \mu e^{\mu(t-1)} \qquad \text{and} \qquad G_X''(t) = \mu^2 e^{\mu(t-1)}$$

$$\Rightarrow \quad G_X'(1) = \mathrm{E}(X) = \mu$$

$$\Rightarrow \quad G_X''(1) = \mathrm{E}(X(X-1)) = \mu^2$$

$$\Rightarrow \quad \mathrm{Var}(X) = \mathrm{E}(X(X-1)) + \mathrm{E}(X) - (\mathrm{E}(X))^2 = \mu^2 + \mu - \mu^2 = \mu$$

Note that the Poisson distribution has the remarkable property that the variance equals the expectation.

The Poisson process

When in the course of time events of a specific type occur, then we are dealing with a so-called Poisson process when the following conditions are satisfied:

- the probability of exactly one event occurring in any very short time interval is approximately proportional to the length of that time interval
- the probability of more than one event occurring in any very short time interval is negligible
- the numbers of events in two non-overlapping time intervals are independent random variables.

When we define $X(t)$ as the number of events occurring in a time interval with length t, then it is possible to prove (using differential equations) that $X(t)$ has a Poisson distribution such that:

$$P(X(t)=k) = e^{-\lambda t}\frac{(\lambda t)^k}{k!}$$

Remark: The conditions above are stated in a rather informal way, but they are sufficient for this course. In Bain & Engelhardt, the conditions are formalised, and a proper proof is given. However, you may skip this, as many of you will as yet be unfamiliar with the essential mathematics to do so.

In the expression above, the parameter μ has been replaced by λt. This makes sense, since it means that the expected value of the number of events is directly proportional to the length of the time interval. The parameter λ is called the **rate** or **intensity** of the Poisson process, and is equal to the mean number of events occurring intyre intervals with a length of 1.

We talked above about time intervals, but Poisson processes also occur in physical space. You will get punctures on the bike - under certain circumstances - according to a Poisson process. The variable t could then be defined as the distance travelled (e.g. in km). But punctures will *not* happen according to a Poisson process when a worn tyre has an increased risk of a puncture (then the condition of proportionality is not satisfied). And if a puncture has been repaired in a sloppy way, the independence will be lost.

The Poisson process makes it clear that the sum of two independent Poisson distributed random variables X and Y is again a Poisson random variable. If X is the number of events in the interval $(0, t\,]$, and Y in the interval $(\,t, t+s\,]$, then $X+Y$ is the number of events in the interval $(0, t+s\,]$. The expectation of the number of events in this interval is $\lambda(t+s) = \lambda\, t + \lambda\, s$, which is the sum of the expectations of the X and Y. This result may indeed very easily be proven by using Theorem 3.11.

Relationship between the binomial and the Poisson distribution

As mentioned at the beginning of this section, we did not yet explain the formula for the Poisson pdf. Now we will focus on the derivation of this formula and at the same time stress the relationship between the Poisson and the binomial distribution.

Assume that the number of people entering a certain shop in one hour (X) is Poisson distributed with $\mu = 6$, so the average number of people per hour is 6. What is the probability that in the next hour 5 people enter? According to the Poisson pdf, this probability is

$$P(X = 5 \mid X \sim \text{POI}(6)) = e^{-6}\frac{(6)^5}{5!} = 0.1606 .$$

Let us try now if we could have found this same probability using the binomial distribution. To that end, we will first divide the interval of one hour into 60 periods of 1 minute each. If we would view each of those 60 minutes as a Bernoulli trial in which either one person enters the shop (a success) or no one enters the shop (a failure), then the binomial distribution would apply with $n = 60$. The probability of success p should be equal to 6/60 ($p = 0.1$), so that the expected number of successes in those 60 minutes (Y_{60}) will again be $\mu = np = 6$. We get as probability according to this binomial model

$$P(Y_{60} = 5 \mid Y_{60} \sim \text{BIN}(60, 0.1)) = \binom{60}{5} 0.1^5 0.9^{55} = 0.1662$$

Although this probability is not very different from the exact Poisson probability of 0.1606, there must be a reason for this difference. And that is because the binomial distribution cannot take into account the possibility that maybe, in some minutes, more than one person will enter.

We will follow the same procedure again, but now in case we divide the period of one hour into 3600 periods of one second each. Now we can approximate X by the binomially distributed random variable Y_{3600} with $n = 3600$ and probability of success p=6/3600=0.00167 (such that the expected value for Y_{3600} will be again equal to 6.

$$P(Y_{3600} = 5 \mid Y_{3600} \sim \text{BIN}(3600, 0.00167)) = \binom{3600}{5} 0.00167^5 0.99833^{595} = 0.1607 .$$

As we can see, this probability is now already very close to the exact probability; the reason is that the probability of more than more person entering in one single second is quite small. The binomial model will therefore become more and more accurate when we divide the period of one hour into more and more very short time intervals. That conclusion can be used to derive the Poisson pdf:

Theorem 3.12

If $X \sim \text{BIN}(n, p)$, and $p = \mu/n$ (with μ a positive constant), then

$$\lim_{n\to\infty} \binom{n}{x} p^x (1-p)^{n-x} = \frac{e^{-\mu}\mu^x}{x!}$$

Proof

$$\lim_{n\to\infty} \binom{n}{x} p^x (1-p)^{n-x} = \lim_{n\to\infty} \binom{n}{x} \left(\frac{\mu}{n}\right)^x \left(1-\frac{\mu}{n}\right)^{n-x}$$

$$= \lim_{n\to\infty} \frac{n!}{x!(n-x)!} \left(\frac{\mu}{n}\right)^x \left(1-\frac{\mu}{n}\right)^{n-x}$$

$$= \lim_{n\to\infty} \frac{\mu^x}{x!} \left(\frac{n}{n}\right)\left(\frac{n-1}{n}\right)\cdots\left(\frac{n-x+1}{n}\right)\left(1-\frac{\mu}{n}\right)^{n} \left(1-\frac{\mu}{n}\right)^{-x}$$

For each fixed value of x, it is clear that:

$$\lim_{n\to\infty}\left(\frac{n}{n}\right)\left(\frac{n-1}{n}\right)\cdots\left(\frac{n-x+1}{n}\right)=1$$

$$\lim_{n\to\infty}\left(1-\frac{\mu}{n}\right)^{-x}=1$$

And a very well-known result from calculus gives:

$$\lim_{n\to\infty}\left(1-\frac{\mu}{n}\right)^{n}=e^{-\mu}$$

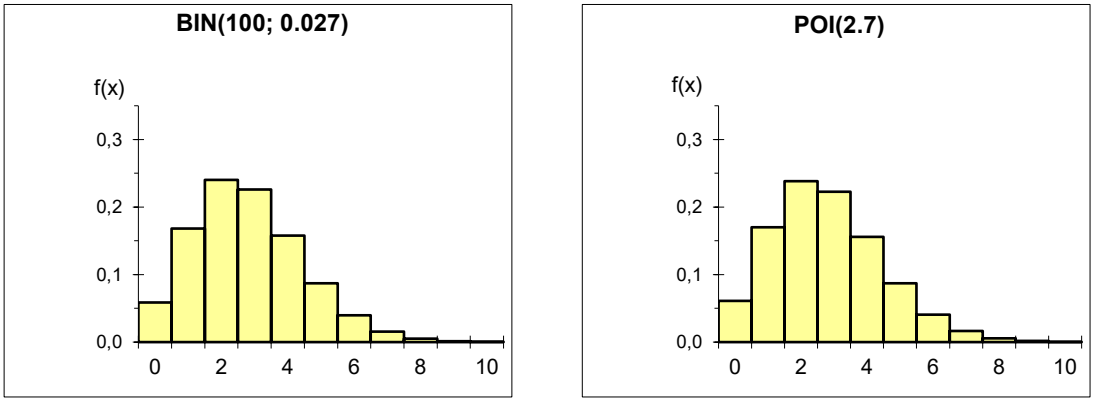
After substitution, we have completed the proof.

But this result also means that, under certain conditions, probabilities from the binomial distribution can be approximated with corresponding probabilities from the Poisson distribution (with $\mu = np$). As a rule of thumb, it is often said that we can do that if $p < 0.05$ and $n \geq 20$. An approximation is also possible if p is very close to 1, say $p > 0.95$ and $n \geq 20$; but instead of counting the number of successes, we should then count the number of 'failures' with probability of failure $q = 1 - p < 0.05$. This *number of failures* will then have by approximation a Poisson distribution (with $\mu = nq$).

Example 3.21

Assume that 99.7% of all medical procedures of a specific type in a hospital are performed without any complications. What is the probability that exactly 499 out of a total of 500 of those procedures will show no complications? This is a binomial probability: $P(X = 499 \mid X \sim \text{BIN}(500, 0.997)) = 0.3349$. In order to approximate this probability using the Poisson distribution, we have to check the rule of thumb. We see that we cannot approximate the number of procedures without complications, since the value of p is far too large. However, we can focus on the number of procedures *with* complications (Y), with the associated probability of 0.003, such that the probability above is also equal to $P(Y = 1 \mid Y \sim \text{BIN}(500, 0.003))$. Since $0.003 < 0.05$ and $500 \geq 20$, we can use the Poisson approximation, with the expected number of procedures with complications equal to $\mu = 500 \cdot 0.003 = 1.5$. The leads to the approximated probability $P(V = 1 \mid V \sim \text{POI}(1.5)) = 0.3347$. ◄

The comparison below shows the pdf for a binomial distribution and the pdf for the associated Poisson approximation.



As we have seen before, the binomial distribution applies to situations with *independent* Bernoulli trials. When n is large and p small, we can use the Poisson distribution as an approximation. But what if there is a very weak dependency between the n Bernoulli trials (weakly dependent)? It appears that the Poisson distribution can still be used very well, where the parameter μ will be set equal to the expected number of successes.

Example 3.22

In exercise 2.25 the birthday problem has been introduced. Now let us see how we can use the Poisson distribution to approximate the probability that in a group with m people, at least two have the same birthday. We will tackle this by comparing the birthdays of just *two* people at a time, which

we will call a trial of the the experiment. For each pair, we are dealing with another trial of the experiment, with a probability of success of 1/365 (the probability of the pair having the same birthday). In total, there will be $\binom{m}{2} = \frac{m(m-1)}{2}$ different pairs (=number of combinations). So we consider $\binom{m}{2}$ trials, each with the same probability of success. However, these trials are not completely independent (why not?), but the degree of dependency will be quite limited. The expected number of pairs with the same birthday will be equal to $\mu = \frac{m(m-1)}{2} \cdot \frac{1}{365}$. This is the parameter we will use in the Poisson approximation.

For example, if $m = 23$: $\mu = \frac{23(23-1)}{2} \cdot \frac{1}{365} = 0.69315$. The desired probability can be approximated by: P (at least two people have the same birthday) = 1 – P(all different birthdays) $\approx 1 - P\,(X = 0 \mid X \sim \text{POI}(0.69315)\,) = 1 - e^{-0.69315} = 0.5000$ (the exact probability is 0.5073).

Incidentally, this approximation provides a useful insight into the reason why this probability increases so quickly as a function of m: the number of combinations of two people (the number of pairs) is already 253 when m = 23! ◄

3.8 Problems

3.1 Dealer A sells 1, 2, 3 or 4 cars per week. The number of cars sold (X) in one week has a probability distribution given by $f(x) = kx^2/x!$ for x = 1, 2, 3, 4.
- a Find the value of k.
- b Find the probability $P(X \geq 2)$, the probability that in a certain week at least two cars are sold.
- c Find the cumulative distribution function (CDF) $F(x)$.

3.2 Find the probability distribution function (pdf) of X if the cumulative distribution function is given by:

$$F(x) = \begin{cases} 0 & \text{for} \quad x < -1 \\ \frac{1}{3} & \text{for} \quad -1 \leq x < 0 \\ \frac{5}{6} & \text{for} \quad 0 \leq x < 2 \\ 1 & \text{for} \quad 2 \leq x \end{cases}$$

3.3 Does there exist a positive value of k for which the function $f(x) = k[1/2^x - 1/4]$ for $x = 0,1,2,3$ could be a pdf?

3.4 A random variable X has a CDF such that: $F(x) = 1 - (1/2)^x$ for $x = 0,1,2,...$ and $F(x) = 0$ for $x < 0$.
- a Find the pdf of X.
- b Suppose that it is also known that X is a discrete variable. Repeat question a.
- c Now, it is also given that X can only assume integer values. Repeat question a.
- d How can we write down the CDF of the random variable in question c?

3.5 A nonnegative integer-valued random variable X has a CDF of the form: $F(x) = 0.1 \cdot (1 + 3x)$ for x = 0, 1, …, k, $F(x) = 0$ for $x < 0$, and $F(x) = 1$ for $x > k$. What is the maximum value that X can attain?

3.6 Suppose that $f_1(x)$ and $f_2(x)$ are both pdf's. Do values of θ exist such that $\theta f_1(x) + (1-\theta) f_2(x)$ is a pdf as well?

3.7 Find the expected value, variance and standard deviation of the random variable X with the following pdf:

x	0	1	2	3	4	5	Total
$P(X = x) = f(x)$	0.02	0.16	0.28	0.33	0.18	0.03	1

3.8 A factory sells per month with probability 0.3 one machine, with probability 0.1 even two machines, but otherwise no machines. The random variable X represents the number of machines sold in a month .
a Find the expected value and the variance of X.
b If the monthly profit (in k€) is equal to 5¼ X − ¾, find then the expected value and the standard deviation of the monthly profit.

3.9 Prove that E($aX + b$) = aE(X) + b where X is a discrete random variable.

3.10 Of 15 electric motors, there are three defective ones. An inspector takes at random three out of these 15 electric motors without replacement, and counts X defective motors. Find E(X) and Var(X).

3.11 A biased coin, with a probability of 2/3 on 'heads', is tossed till 'heads' is tossed or till the number of tosses is five. If X is the number of tosses, find then E(X) and Var(X).

3.12 For which value of p does the variance of the discrete random variable X attain its maximum, if $P(X=0)=P(X=2)=p$ and $P(X=1)=1-2p$ (with $0 \le p \le 1/2$).

3.13 In a lot of six batteries one is worn out. A mechanic tests the batteries one by one until a faulty battery is found. After testing, he puts the already tested batteries aside, but after every three tests he takes a break, and during each break a silly fellow puts one of the three tested batteries back with the other untested batteries.
a Find the probability function of X, the total number of tests needed until the faulty one is found.
b Assume that the first test of each set of three costs € 5, and each of the other two tests € 2. Find the expected value of the *extra* costs that are caused by the behaviour of the silly fellow.

3.14 The discrete random variables X_1 and X_2 have respectively the pdf's $f_1(x)$ and $f_2(x)$. Let Y be the random variable with pdf $f_Y(y)=\theta f_1(y)+(1-\theta)f_2(y)$ for $0 \le \theta \le 1$.
a Express the expected value and the variance of Y as a function of the expected values and variances of X_1 and X_2 .
b Is the random variable Y the same as $\theta X_1 + (1-\theta)X_2$?

3.15 The St.Petersburg paradox was described by Nicolaus Bernoulli, professor of St. Petersburg (his uncle was Jacob Bernoulli, whose name is linked to the famous Bernoulli distribution). Someone pays N euros to join a game at the casino. An unbiased coin is getting tossed . The casino will pay 2 euros as ' heads' appears at the first toss, 4 euros as ' heads ' appears for the first time at the second toss and 2^i euros as 'heads ' appears for the first time at the i-th toss..
a What is the expected profit per game for the casino?
b If the previous question is answered correctly, then it follows that the casino should never allow this game, regardless of the size of N. However, it was played in that time. Can you guess why?
c Suppose the casino paid out a sum 2^i if $i \le 19$ and otherwise 2^{20} (without going bankrupt). Calculate the height of the initial payment N that should be paid to make the game fair (meaning that the expected profit per game is 0).

3.16 Let X be a nonnegative integer-valued random variable, so possible outcomes are 0, 1, 2, ... Find a relationship between $P(X \ge 0) + P(X \ge 1) + P(X \ge 2) + \ldots$ and E(X).

3.17 Suppose that the average amount of cash students carry with them is equal to 16 euros. What is the upper limit for the probability that a random student has 100 euros or more in the pocket ?

3.18 Prove Theorem 3.6 by using Theorem 3.5.

3.19 For a discrete random variable X the following pdf is given:
$f(x)=1/8$ (for $x=4,6$) and $f(x)=6/8$ (for $x=5$)
a Find the expected value μ and standard deviation σ.
b Compute $P(|X-\mu| \ge k\sigma)$ for $k=2$. Compare this result with the upper limit that follows from Chebyshev's inequality.

3.20 Bolts are packed in boxes, such that the average number of bolts in each box is 100, with a the standard deviation equal to 3. Use Chebyshev 's inequality to determine a limit to the probability that the number of bolts in an arbitrary box is at least 95 and at most 105.

3.21 Consider a random variable X with $P(X=1)=0.5$ and $P(X=2)=0.5$. Show that the inequality of Jensen holds in this case, when we define the function $g(x)=1/x$.

3.22 Show in two different ways that for an arbitrary random variable X it is true that: $(\mathrm{E}(X))^2 \le \mathrm{E}(X^2)$.

3.23 Consider an arbitrary random variable X, and define $Y = X^3$.
- a Can we say something about the relationship between $\mathrm{E}(Y)$ and $(\mathrm{E}(X))^3$ by using the inequality of Jensen?
- b Now it is given, as additional information, that the support consists only of positive values for x. Answer again the question in part a given this additional information.
- c Now it is given, as additional information to part a, that the support consists only of negative values for x. Answer again the question in part a given this additional information.

3.24 $G_X(t) = (t + 1)^5 (t + 2)^5 / 6^5$ is the probability generating function of a random variable X.
- a Find $\mathrm{E}(X)$.
- b What are the possible outcomes for X?

3.25 Prove that for a discrete random variable X with probability generating function $G_X(t)$ it is true that: $P(X = i) = G_X^{(i)}(0) / i!$.

3.26 Let X be a random variable with the following probability generating function: $G_X(t) = c(1 + t)^3$
- a Find the value of c.
- b Find $P(X = 1)$.
- c Give the expected value of X.

3.27 The probability generating function of X is $G_X(t) = (0.5 + 0.4t)^2 + 0.19t^3$.
- a Find $P(X \ge 2)$.
- b Find the expected value and variance of X from $G_X(t)$.

3.28 Give the probability generating function of
- a the number of dots in one throw of a die .
- b the number of 'heads' if an unbiased coin is tossed twice.
- c the number of 'heads' if an unbiased coin is tossed three times.
- d the number of 'heads' if an unbiased coin is tossed n times.

3.29 Find the probability generating function of a uniform distribution on $\{2, 3, 4, 5\}$.

3.30 Find the expected value and variance of the discrete uniform distribution on [a, b].

3.31 Give a formula for the cumulative distribution function of X if $X \sim$ DU(a, b).

3.32 If $X \sim$ DU(a, b) and $Y \sim$ DU(2a, 2b), can Y then be described as a linear function of X ?

3.33 The distribution of a discrete uniform variable X has expected value ½ and variance ¼. Prove that X has a Bernoulli-distribution.

3.34 Assume $X \sim$ BIN(10, 0.3). Write down the probability generating function, and determine, by *using this probability generating function*, the probabilities $P(X = 0)$, $P(X = 1)$ and $P(X = 2)$.

3.35
- a Find the probability of three successes if $p = 0.25$ and $n = 15$ (using the binomial distribution).
- b In a multiple choice test consisting of fifteen questions, each with four possible answers, someone answers each question just by lucky guess. Calculate the probability that at least four questions and at most seven questions are answered correctly.

3.36
- a Find the probability of at least ten successes if $p = 0.7$ and $n = 15$.
- b If someone has at a multiple choice exam with fifteen questions for each question the same probability p of giving the correct answer, find the smallest value of p such that the probability of answering at least ten questions correct is at least equal to 0.8? (You have to choose from the possibilities that are given by the table , or which can be derived from this table).

3.37 Men do experiments with different concentrations of chemicals to test a pesticide against dandelions. At a certain concentration the probability is 80% that the dandelion is destroyed within a day. Twenty dandelions are exposed to this concentration. You may assume that the dandelions respond independently of each other to this concentration of chemicals.
- a Find the probability that exactly 14 dandelions are destroyed.
- b Find the probability that at least 10 dandelions are destroyed.

3.38 Find the probability that if an unbiased die is thrown 240 times the total number of twos and threes is at least 75, but not more than 83. You can use a computer or a calculator.

3.39 A red die is unbiased, a green die has a probability 1/10 on throwing six.
a Find the probability of three sixes if the red die is thrown three times.
b Find the probability of at least 30 sixes in 100 throws with the red die. You can use a computer or a calculator.
c Find the probability of exactly 3 sixes in total if the green die is thrown five times and the red die is thrown four times.

3.40 You toss in three rounds with the same four unbiased coins. Find the probability that in exactly two of those rounds 'heads' will appear exactly one time.

3.41 With a probability of 0.2 a baby lets his parents sleep through the night, regardless of the day of the week, and regardless of what happened in previous nights. Let X be the number of times in a week that the parents can sleep through the night.
a Find the probability distribution function of X.
b Find the probability that the parents can sleep through the night at least five times in a week.
c Find the conditional probability $P(X \geq 5 \mid X \geq 3)$.

3.42 20% of the descendants of a particular plant have characteristic K.
a Find the probability that at most 13 out of a random sample of 100 descendants of this plant have characteristic K.
b Find the probability that in two groups of 100 descendants of this plant in total exactly 26 descendants have characteristic K. You can use a computer or a calculator.

3.43 For every booked seat of a regular service with an aircraft the probability is 10% that the passenger does not show up. The aircraft has 90 seats. Sometimes more than 90 seats are booked. (You can use a computer).
a Find the probability that the aircraft has enough seats for the passengers that show up if 95 seats are booked.
b Compute for which number of booked seats the probability that the aircraft has not enough seats for the passengers that show up is at most 1%?

3.44 A blue die shows the image of a skull on three of the six faces, and a yellow die on four of the six faces. The blue die is rolled four times and the yellow die is rolled twice. Find the probability of a total of five skulls.

3.45 A faculty wants to have 360 freshmen left over after one semester. Experience shows that on average 15% do not 'survive' the first semester. If the faculty could decide how many students may start the first semester, calculate then how many students should be accepted at the start of the year in order to keep at least 360 freshmen with probability 0.5 after one semester. (Assume that the 'survival'-probability is the same for each freshman, independent of the survival of the other freshmen. You can use a computer).

3.46 The number of cars sold in one week at dealer A has a BIN(9, 0.2)-distribution.
a Find the expected value and standard deviation of the number of cars sold in one week at dealer A.
b The fixed costs are € 1000 each week; per car sold the earnings are € 1500. Find the expected value and standard deviation of the net weekly income.

3.47 A bicycle shop sells two types of bikes. Bicycle B provides € 110 profit per bike and bicycle C is good for € 140 profit per bike.
Weekly, X bicycles of type B, $X \sim$ BIN (5, 0.2), and Y bicycles of type C, $Y \sim$ BIN (8, 0.1), are sold, independently of each other.
a Find the expected value and standard deviation of the number of bicycles sold in one week.
b Find the expected value and standard deviation of the weekly income.

3.48 If X and Y are two independent binomially distributed variables with parameters n, p_1 and m , p_2 respectively (with $p_1 \neq p_2$), then the sum $(X + Y)$ is not binomially distributed. Prove this proposition by using probability generating functions.

3.49 The "drunkard's walk" problem . This problem is named after the drunkard who has just left the pub, and takes a step to the right with probability ½ , or a step to the left with probability ½ . Each step is assumed to be independent of the other steps. This model, of which also two- and three-dimensional versions exist, has surprisingly serious applications in physics, chemistry, biology etc.

Define D_m as the distance to the starting point of the walk after *m* steps. (D_m is always greater than or equal to zero, because we are not interested in the direction, but only in the distance).

a Find $\mathrm{E}(D_1)$, $\mathrm{E}(D_1^2)$, $\mathrm{E}(D_2)$, $\mathrm{E}(D_2^2)$, $\mathrm{E}(D_3)$, $\mathrm{E}(D_3^2)$

b Now we are going to tackle this problem in a more general way. Let *m* be the total number of steps that is made at some stage, and define the random variable *X* as the number of steps (of those *m* steps) that is done to the right. So, $X = 0, 1, 2, \ldots, m$.
Find $P(X = x)$.

c Now define $D_m^2(X)$ as the square of the distance to the starting point of the walk after *m* steps as a function of *X*. Show that $D_m^2(X = x) = (m - 2x)^2$

d Show that $\mathrm{E}(D_6^2(X)) = 6$. (Use $\mathrm{E}(g(X)) = \sum_x g(x) \cdot P(X = x)$)

3.50 You fight in a championship against an opponent who is stronger than you are. There are 2, 4 or 6 matches played, and the one who wins more than half of the matches is the champion (if both win half of the matches, then there is no champion). Assume that the probability that you win a match is equal to 5/12. The results of the various matches are independent of each other. Because you are weaker than the opponent, he/she gives you the choice how many games you want to play (i.e. 2, 4 or 6). If your goal is to be the champion, what is the best choice? Answer this question first intuitively, and then find the probabilities of becoming the champion.

3.51 A bin contains 50 fuses, including seven broken ones. Ten fuses are drawn from this bin without replacement. Find the probability that the sample contains at least one broken fuse.

3.52 A batch of 400 tyres, including 10 damaged ones, will be checked before shipment. A sample of five tyres will be drawn without replacement. The parts a, b and c below describe three different situations.

a If the sample does not include damaged tyres, the whole batch will be sent. Find the probability that the batch will be sent.

b Let *X* be the number of damaged tyres in the sample. lf this sample does contain one (or even more) damaged tyres, then these are replaced by tyres that are not damaged, after which the batch will be sent. Find the expected value of *X* and use this expected value in order to find the expected value of the fraction of damaged tyres in the batch that will be sent.

c If the sample of 5 tyres contains a damaged one, all 400 tyres will be checked and all damaged tyres will be replaced by ones that are not damaged. Find again in this case the expected value of the fraction of damaged tyres in the batch that will be sent.

3.53 A sample of size 4 is drawn without replacement from a batch of 25 pieces, of which *D* are defective. Find the probability that all 4 pieces drawn are not defective as a function of *D*.

3.54 Show the following statement by expanding: $\dfrac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} = \dfrac{\binom{n}{x}\binom{N-n}{M-x}}{\binom{N}{M}}$. Give an explanation of this equality by interpreting the formulas.

3.55 A box contains six red and four green balls. These balls are successively drawn without replacement in a random order.

a Find the probability that the third green ball is drawn at the fifth drawing.

b Which drawing is the most probable one for drawing the third green ball?

3.56 A bag contains three red and five blue marbles. These marbles are drawn consecutively, so without replacement, from the bag. Let *X* represent the number of the drawing at which the last red marble is drawn.

a Find the probability distribution function of *X*.

b Find E(*X*) and Var(*X*).

3.57 Prove that the HYP(n, M, N)-distribution converges to the binomial distribution BIN(n,p) if we let

$N \to \infty$ and $M \to \infty$, in such a way that $M/N = p$ remains constant. So, prove that:

$$\lim_{N\to\infty} \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} = \binom{n}{x} p^x (1-p)^{n-x}$$

3.58 Assume $X \sim$ BIN(80, ¼). Find $P(X = 15)$ with the aid of a computer, and compare this probability with successively $P(X = 15 \mid X \sim \text{HYP}(80, 25, 100))$, $P(X = 15 \mid X \sim \text{HYP}(80, 50, 200))$, $P(X = 15 \mid X \sim \text{HYP}(80, 250, 1000))$ and $P(X = 15 \mid X \sim \text{HYP}(80, 2500, 10000))$.

3.59 Let X be the number of 'heads' that player A tosses with two unbiased coins, and Y the number of 'heads' that player B tosses with three unbiased coins. If $X > Y$, then player A wins; if $X < Y$, then player B wins; if $X = Y$, then the game will be replayed.

a Find the probability that player A wins in a single turn (so in the first play).

b Find the probability that player A wins.

3.60 From a large group of applicants 80% is qualified for a particular function. Interviews are held with the candidates - in random order - and the first one that is qualified will get the job. Find the probability that at most four such interviews are required.

3.61 The probability that a group of lions on a hunt will catch the prey is 0.4. Find the probability that the group of lions needs in order to catch a prey:

a exactly 6 attempts..

b less than 6 attempts.

c at least 5 attempts.

3.62 Use the *binomial* distribution for calculating the probability in part b of the previous problem. Remember that "less than 6 attempts " means that the group in five consecutive attempts (real or virtual) at least one time must have been successful.

3.63 The random variable X is the number of times that an unbiased coin is tossed until the first time that 'heads' appears. Find the probability that X is even.

3.64 The 'coupon collector's' problem. It asks the following question: Suppose that there is an urn of n different coupons, from which coupons are being collected, equally likely, with replacement. Given n coupons, how many coupons do you expect you need to draw with replacement before having drawn each coupon at least once? This problem is named after the collector of images ("collect all 160 baseball stars") that are sometimes included in the packaging of products. If we assume each package to contain a single image, and if each different image is equally likely to be present, then the question is how many of the product has to be purchased (expected) until all images have been collected.
Consider first the well-known die. What is the expected value of the number of throws of the die until each number (from 1 till 6) comes up at least once. Define the random variable X as the number of throws. Now, also define the random variables Y_i (i = 1, 2, 3, 4, 5, 6) as the number of throws required to go from (i -1) different numbers to i different numbers. Of course $P(Y_1 = 1) = 1$, because after one throw we have in any case the first number (1, 2 , 3, 4 , 5 or 6) . Y_2 is then the number of (extra) throws required until the second (distinct) number comes up.

a Which kind of distribution has Y_2, and what is the corresponding expected value?

b What is the expected value of Y_i (i = 3, 4, 5, 6)?

c Use the general rule $E(Z+Y) = E(Z) + E(Y)$ to find $E(X)$.

d Now the general case. Someone wants to collect n images. Each packing contains one image. Give a formula for the expected value of the number of packings that this person has to buy before he possesses all the images.

A mathematical approximation formula states that $1 + \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{n} \approx \ln(n) + \gamma + \frac{1}{2n}$, in which γ is the constant of Euler (= 0.57722...).

e What is the expected value of the number of people that is required before we have gathered a group of people in which each of the 365 possible birthday dates is represented?

3.65 Assume $X \sim$ NEGBIN(n, p). Find the expected value and the variance of X by using the probability generating function of X.

3.66 When throwing an unbiased die, a result of 3 dots or more is defined as a success.
- a Find the probability that the first success occurs at the third throw.
- b Find the probability that an even number of throws is needed up to and including the moment that the first success occurs.
- c One throws with this die until the fifth 'success' has occurred. Find the probability that this occurs within at most nine throws.
- d Find the probability that the number of attempts up to and including the first failure (i.e. 1 or 2 dots) is divisible by 4.

3.67 To take an exam you must pass nine tests. Assume that any attempt has a probability of success of 60%.
- a Find the probability that exactly at the 15th test the examination requirement is met.
- b Use a binomial table in order to find the probability that only áfter the 15th exam attempt the examination requirement is met (so, at least 16 attempts are needed).

3.68 An unbiased coin is tossed. Find the probability that 'tails' is thrown at most twice before the fourth time 'heads' occurs.

3.69 An unbiased die is thrown and the number of fives and sixes are counted.
Use in answering the questions below the additional, extended binomial tables that can be found on the blackboard-site.
- a Find the probability that the second six will occur at the latest within 11 throws.
- b Find the probability that the sixth 5 or 6 will occur after the 20^{th} throw.

3.70 Suppose that $X \sim$ NEGBIN(r, p). Find the expected value and variance of the variable $Y = X - r$, which means the number of failures preceding the r-th success ['the other definition'].

3.71 Find also the probability generating function of the variable $Y = X - r$, being the number of failures preceding the r-th success.

3.72 According to a rule of thumb the interval $(\mu - 2\sigma\ ;\ \mu + 2\sigma)$ contains about 95% (sometimes more) of a probability distribution (if the distribution is approximately symmetrical and uni-modal).
- a Find this interval for the number of throws that is necessary in order to get the tenth six with an unbiased die.
- b Find with the inequality of Chebyshev the minimum probability for the interval found.

3.73 For which value of p is $P(X = t)$ the highest, if $X \sim$ NEGBIN(r, p)?

3.74 Prove that a Negative Binomial distribution cannot be symmetrical.

3.75 In a box of 25 old light bulbs five of them are defective, because of rough handling. Successively light bulbs are drawn out of this box. Find the probability that the third and the fourth defective light bulb are found in the fifth resp. sixth draw if
- a each light bulb is replaced into the box after a check.
- b the bulbs are successively discarded and not replaced.

3.76 Suppose that $X \sim$ POI(μ). Find the probability of exactly three events as $\mu = 1.5$ by using the table and compare your answer with the probability that is obtained if you use the probability density function.

3.77 Find the probability of at least three and at most six lightning strikes in the month of July in town Z if the average number of lightning strikes in the month of July is 10.

3.78 Suppose that $X \sim$ POI(a) and $Y \sim$ POI(b), with X and Y independent variables. Find the probability density function of $X + Y$. Do this:
- a by using the probability generating function.
- b by using the convolution formula (see section 3.6).

3.79 Traffic accidents at a particular intersection occur according to a Poisson distribution with an expectation of 4 per year. You may assume that in all periods of the year the probability of an accident at this intersection is the same.
- a Find the probability of at most one accident in one specific year at this intersection.
- b Find the probability of exactly three accidents in half a year.
- c At another intersection on average two accidents happen per year. Find the probability of at least three accidents in one year at both intersections taken together.

3.80 Let X be the number of hits in a baseball game and assume that $X \sim$ POI(α). It is known that the probability of no hits, $P(X = 0)$, is equal to 1/3. Find α.

3.81 In one supermarket each morning on average three men and five women (independent from each other) enter per minute. Find the probability that at least 20 customers come in on a certain day between 10:20 and 10:25 hours.

3.82 In a specific police station serious crimes are reported with a frequency of on average five per night. Assume the the number of serious crimes reported has a Poisson distribution.
a Find the probability of at most three serious crimes reported in a given night.
b Find the probability that a full week passes by in which every night no more than three serious crimes are reported.

3.83 Deaths in a given town follow a Poisson distribution with an average of five per week.
a Find the expected value of the number of deaths in a period of three days.
b Find the probability that anyone survives a certain three-day period.
c Describe the probability that at least 250 people die in one year, by using the cumulative distribution function F.

3.84 A bakery bakes 1000 cookies with pieces of chocolate in it. Available are a total of n pieces of chocolate, that are thoroughly mixed into the dough, which will result in the number of pieces of chocolate per cookie having (approximately) a Poisson distribution .
a If $n = 4900$, then find which proportion of the cookies will contain at least two pieces of chocolate.
b If $n = 4900$, then find the expected number of cookies with exactly three pieces of chocolate.
c If more than 1% of the 'cookies with pieces of chocolate ' do not contain a single piece of chocolate, then problems arise with the 'inspection bureau of sweets' from the government. Which value of n should at least be taken in order to ensure that the probability that a cookie contains no chocolate, is at most equal to 1%?

3.85 Suppose that the probability (distribution) function of a random variable X is given by:

$$f(x) = \frac{1}{(e-1)(x!)} \text{ for } x \in \mathbb{N} = \{1, 2, 3, \ldots\}$$

Find the probability generating function and calculate with the aid of this function the expected value and variance of X.

3.86 (B&E, 3.28) Suppose that $X \sim \text{POI}(10)$.
a Find $P(5 < X < 15)$.
b Use Chebyshev's inequality to find a lower bound for $P(5 < X < 15)$.
c Find a lower bound for $P(1 - k < X/10 < 1 + k)$ for arbitrary $k > 0$.

3.87 At a certain police station suspects are arrested according to a Poisson process with an average of two 'catches' a day. The cell capacity is 3. If there are more than three suspects brought in on one day, then the excess is sent away.
a Find the probability of exactly three arrests per day.
b Find the probability that suspects are sent away on a certain day.
c Find the probability (distribution) function of X, the actual number of jailed suspects per day.
d Find the expected value of the number of suspects that are sent away per day.
e It is decided to build extra cells, such that a sufficient cell capacity is present on approximately 95% of the days. How many extra cells should be built?

3.88 Telephone calls arrive at a telephone exchange with a frequency of on average 6 per minute and follow a Poisson process. The probability that an incoming call is connected correctly is 0.8 (independent of the frequency and earlier failure). Show that the number of correct calls connected in a minute follows a Poisson distribution with $\mu = 4.8$.

3.89 In one liter of dough on average 200 raisins are mixed well. If one makes rolls of raisins with this dough, in which 125 cm^3 is used, then calculate
a the probability of exactly 22 raisins in one roll.
b the probability of in total exactly 52 raisins in two rolls.

3.90 Three unbiased dice are thrown 500 times. Find the probability of twice the event 'three sixes', using both the exact distribution as well as using a Poisson approximation. (Use the formulas and a calculator.)

3.91 In a production process an average of 99 % of the items is without manufacturing defects. Find the probability that a retailer, who receives 300 of these items, will get at least 295 items without manufacturing defects, assuming that the defective ones will appear at random, so without dependency, clusters etc.

3.92 20% of the chips produced in a particular production line contains too much fat. A random sample of 100 chips is taken. Find the probability that at most 13 are too fat
   a exactly, with the use of a binomial distribution. (use a calculator)
   b by using an approximation, although $p = 0.2$ is actually too high, with a Poisson distribution. (use a calculator)

3.93 You live in Amsterdam , a city with 700,000 inhabitants. You meet someone for the first time, who also comes from Amsterdam. Both of you know 500 inhabitants of Amsterdam, and assume that the next (improbable) property holds: the friends of both of you are to be regarded as a random sample of the inhabitants of Amsterdam (i.e. the probability that you know any welfare mother, is the same size as the probability that you know a random student) .
   a What is the probability that you will have at least one acquaintance in common?
   b Find the same probability by using a Poisson approximation.

3.94 On a scratch card used in a scratch lottery there is a readable number (between 1 and n), and a number below the scratch layer (also between 1 and n). There are no two scratch cards with the same readable number, and also no two scratch cards with the same unreadable number. Assume that the unreadable numbers are randomly printed on the scratch cards.
   a The lottery sells all n scratch cards. Use a Poisson approximation in order to approximate the probability that no one wins the lottery (so not one of the n readable numbers did match with the unreadable numbers).
   b The lottery sells half of the scratch cards. Find the probability that exactly two people win the lottery.

3.95 Compare with Example 3.22: Problems that are hard to solve in an exact way, can often be approximated in a good way by using the Poisson distribution. A good example is the near-birthday problem: What is the probability that in a group of m people at least two will have their birthday within two days (so either on the same day or with a difference of one day). Determine the probability of this event for $m = 20$.

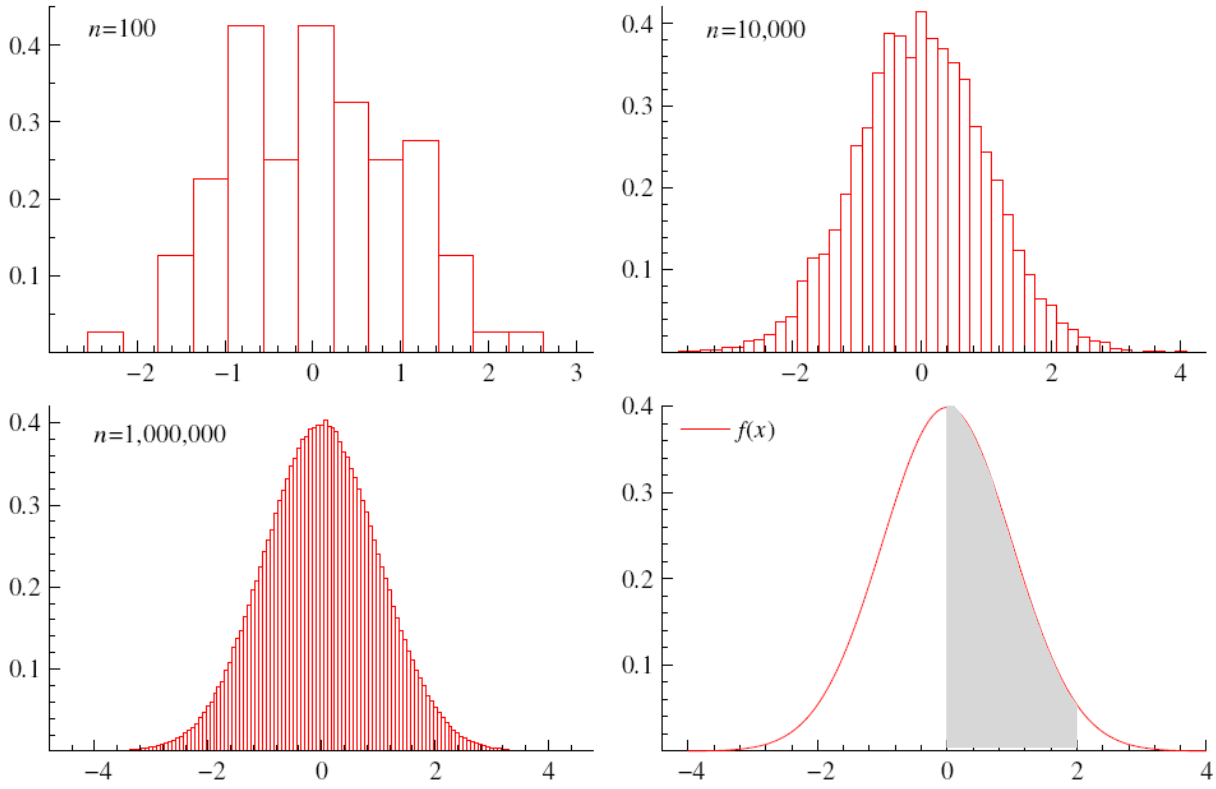
4 Continuous random variables

4.1 Introduction

In the previous chapter we discussed discrete random variables, so random variables which can only assume a countable number of possible values. In chapter 1 we have already seen that the data can also be *continuous* in nature; similarly, random variables are called continuous when they can take on any real value within one (or more) interval(s). This is the case, for example, when we measure the weight X of any person with an infinite accuracy. What is the probability that a specific person has a weight of *exactly* 60 kg, so 60.000000000000000000000000000000… kg? Not surprisingly, this probability will be infinitely small, in other words it will be equal to 0. This applies, in general, whenever X is continuous: $P(X = x) = 0$ for all x! This also means that the way in which the probability distribution function has been defined in Chapter 3 (see Definition 3.1) is no longer useful for continuous random variables. On the other hand, the definition of the cumulative distribution function $F(x) = P(X \leq x)$ (see Definition 3.2) remains to be very useful.

To replace the probability *distribution* function, we will use the probability *density* function for continuous random variables. We encountered the concept of density already at page 6 when we were discussing histograms. Suppose we take a sample of 100 observations from a continuous distribution, and we use this sample to draw a density histogram (i.e. on the vertical axis the density is displayed where density = relative frequency / column width). The result could, for example, look like the first figure below. The total surface area of all the columns added together should be equal to 1 by definition (check!). If we increase the number of observations in the sample to 10,000 and then to 1,000,000, we might find histograms as in the second and third figure below. Of course, the total surface area remains 1.

If we continue to increase the number of observations, we can see that the shape of the density histogram will converge to the graph shown in the fourth figure; it shows the curve of the so-called density function, for which the total area under the curve (and above the horizontal axis) will always be 1. In fact, surface areas under this curve can be interpreted as probabilities, which follows directly from the concept of probability as relative frequencies (see section 2.4.2). So the surface for the grey area in the last figure represents the probability $P(0 \leq X \leq 2)$.



4.2 Probability density function and cumulative distribution function

(B&E, Page 62-66)
The definition for the CDF in chapter 3 (see Definition 3.2), which is still valid for continuous random variables, implies that each CDF is a nondecreasing function. We saw that the CDF for a discrete random variable shows a jump upwards at each possible outcome for the random variable, and the size of the jump equals the probability of the outcome occurring. The CDF was therefore *discontinuous* in all those points. But when the CDF is a continuous function, then we call the associated random variable *continuous*.

Definition 4.1 *(B&E, Def.2.3.1)*

A random variable X is called a **continuous random variable** if a function $f(\cdot)$ exists such that the cumulative distribution function (CDF) $F(x)$ can be written as:

$$F(x) = P(X \le x) = \int_{-\infty}^{x} f(t)dt \, .$$

The function $f(\cdot)$ is called the **probability density function (pdf)** (Dutch: kansdichtheid of (kans)dichtheidsfunctie).

Note that the above definition implies that the CDF for continuous random variables does not have any discontinuity points. In intervals where the CDF is *strictly* increasing, any real number is a possible outcome for the random variable. However, because no jump upwards occurs at any point (because the CDF is continuous), we must draw the conclusion that the probability of any specific value to occur is equal to 0, so $P(X = x) = 0$ for any x. Only probabilities of an outcome to fall *within an interval of values* can have values greater than 0.

Probabilities can be found by determining the relevant surface area under the pdf. For $a < b$:

$$P(a < X \le b) = P(X \le b) - P(X \le a) = F(b) - F(a) = \int_{-\infty}^{b} f(x)dx - \int_{-\infty}^{a} f(x)dx = \int_{a}^{b} f(x)dx \, .$$

Note that $P(a < X \le b) = P(a \le X \le b) = P(a \le X < b) = P(a < X < b)$!

Calculus shows that the pdf can be found directly by taking the derivative of the CDF:

$$f(x) = \frac{dF(x)}{dx} = F'(x)$$

Because the CDF is by definition a nondecreasing function, the pdf tells us how strongly the CDF increases in value when the variable x increases marginally, because:

$$f(x) = \frac{dF(x)}{dx} = \lim_{h \downarrow 0} \frac{F(x+h) - F(x)}{h} = \lim_{h \downarrow 0} \frac{P(x < X \le x+h)}{h}$$

Remark: often the CDF will not be differentiable for every value of x as can be seen for example in Example 4.1 at $x = -1$ and $x = 1$. However, this will only happen in general at a finite set of singular points, and will therefore cause no problems; keep in mind that the value for an integral like in Definition 4.1 will not change by arbitrarily changing the integrand at a finite number of values.

Theorem 4.1 *(B&E, Th.2.3.1)*

Each pdf $f(x)$ for a continuous random variable X has the following properties:

1. $f(x) \ge 0$ (for $x \in \mathbb{R}$), and 2. $\int_{-\infty}^{\infty} f(x)dx = 1$

Proof

The first property follows from Definition 4.1 considering the fact that each CDF is a nondecreasing function by definition. For the second property, we have:

$$\int_{-\infty}^{\infty} f(x)dx = F(\infty) = P(X \le \infty) = 1 .$$

The set of all possible values for the random variable, which is equal to the set of all values x for which $f(x) > 0$, is again called the *support*. Usually, we will only define a pdf by specifying its values for all x belonging to the support set, thereby implying that $f(x) = 0$ for all other values of x.

Note that we used the same abbreviation (i.e. 'pdf') for the probability density function in Definition 4.1 as we did for the probability distribution function for discrete random variables. This is often convenient in future discussions, but one must always be aware of the differences. Fortunately, just a look at the formulation of the support set for a specific pdf will tell us the type we are dealing with.

Example 4.1

Say $f(x) = 0.6 - 0.3x^2$ for $-1 \le x \le 1$ (so $f(x) = 0$ for x otherwise; X is continuous)

It is easy to check that the pdf on the interval (–1, 1) is not negative, and that the integral is 1:

$$\int_{-\infty}^{\infty} f(x)\,dx = \int_{-1}^{1} (0.6 - 0.3x^2)\,dx = \Big[0.6x - 0.1x^3\Big]_{-1}^{1} = 1.2 - 0.2 = 1$$

Note the first step above: since $f(x) = 0$ for $x < -1$ and for $x > 1$, we must make sure to use the correct integration limits when we replace $f(x)$ by $0.6 - 0.3x^2$.

Say we need to determine $P(-0.2 < X < 0.6)$. Then we get:

$$P(-0.2 < X < 0.6) = \int_{-0.2}^{0.6} (0.6 - 0.3x^2)\,dx = \Big[0.6x - 0.1x^3\Big]_{-0.2}^{0.6} =$$

$$= 0.3384 - (-0.1192) = 0.4576$$

The CDF (for $-1 < x < 1$) can now be determined by:

$$F(x) = \int_{-\infty}^{x} f(u)\,du = \int_{-\infty}^{-1} 0\,du + \int_{-1}^{x} (0.6u - 0.3u^2)\,du =$$

$$= \Big[0.6u - 0.1u^3\Big]_{-1}^{x} = 0.6x - 0.1x^3 + 0.5 \quad \text{on } [-1, 1]$$

For $x \le -1$ we clearly get $F(x) = P(X \le x) = 0$, and $F(x) = 1$ for $x \ge 1$.

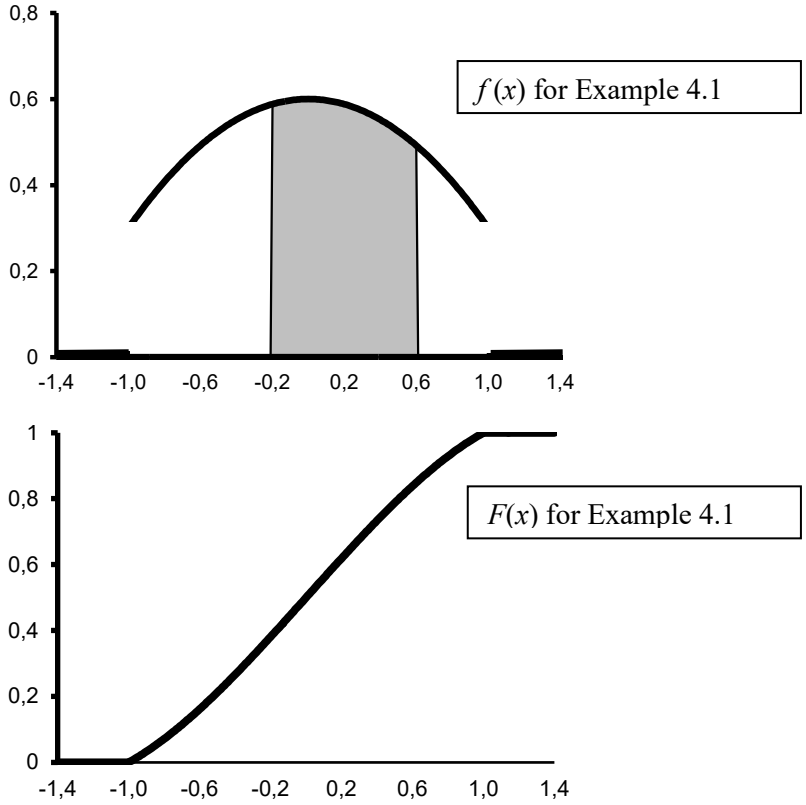
The other way around, we could have determined the pdf simply by taking the derivative of the CDF, resulting in

$$f(x) = F'(x) = \begin{cases} 0.6 - 0.3x^2 & \text{for } -1 < x < 1 \\ 0 & \text{for } x < -1 \text{ and for } x > 1 \end{cases}$$

Note that the pdf cannot be determined at $x = -1$ and $x = 1$ by taking the derivative, because these points are in this example discontinuity points for the pdf; however, it really does not matter at all, and we are from a mathematical point of view completely at liberty to give the pdf any value we would like at any collection of singular points. Usually, we will just write either $-1 \le x \le 1$ in the first row above, or $x \le -1$ and $x \ge 1$ in the second row.

When the CDF is already known, then the probability $P(-0.2 < X < 0.6)$ (see above) could also have been found as follows:

$P(-0.2 < X < 0.6) = F(0.6) - F(-0.2) = 0.8384 - 0.3808 = 0.4576.$ ◄



Many novices to probability theory make the persistent error by trying to determine the CDF as an *indefinite* integral (i.e. an integral without limits) instead of a definite integral. Very often, this will lead to incorrect results (you can try it for the example above)! You can check the resulting CDF for errors like these, because we know that $F(x)$ must be 0 for x equal to the smallest possible outcome, and $F(x)$ must be 1 for x equal to the largest possible outcome (for the random variable).

Remark: random variables may also have a mixed distribution, that is to say they are not purely continuous or discrete, but have a mixed form. For example, suppose the random variable X is defined as the time a person has to wait at a traffic light. It is quite possible that there is no need to wait at all, so $P(X = 0) > 0$, which means that the distribution is not purely continuous. But, in case the person does have to wait, then that waiting time will be continuous. Again, the CDF can be used without a problem, but working with the pdf becomes a bit tricky because of the difference in interpretation between a discrete and a continuous probability function. We can also work with conditional distributions (discussed further in course Probability Theory and Statistics 2), for example the distribution of the waiting time *given* that the person has to wait. This latter distribution is again a purely continuous distribution.

4.3 Expected value, variance, inequalities

(B&E, Page 67-77)
With some minor changes (integration instead of summation) the results of sections 3.3 and 3.4 are also valid for continuous random variables. We will recall those here (without proofs).

Definition 4.2 *(B&E, Def.2.3.2)*

The **expected value** or expectation or mean (Dutch: verwachte waarde of verwachting), notation: $\mathrm{E}(X)$ or $\mathrm{E}X$ or μ), of a continuous random variable is defined by:

$$\mu = \mathrm{E}(X) = \int_{-\infty}^{\infty} x\, f(x)\, dx$$

if the integral converges absolutely; otherwise, we say that $\mathrm{E}(X)$ does not exist.

Theorem 4.2

If X and Y are random variables, $g(\cdot)$, $g_1(\cdot)$, $g_2(\cdot)$ are real functions (on a domain including the support of X) and a, b, and c are constants, then:

$$\mathrm{E}(g(X)) = \int_{-\infty}^{\infty} g(x)\, f(x)\, dx$$
$$\mathrm{E}(c) = c$$
$$\mathrm{E}(cg(X)) = c\mathrm{E}(g(X))$$
$$\mathrm{E}(g_1(X) + g_2(X)) = \mathrm{E}(g_1(X)) + \mathrm{E}(g_2(X))$$
$$\mathrm{E}(aX + b) = a\mathrm{E}(X) + b$$
$$\mathrm{E}(X + Y) = \mathrm{E}(X) + \mathrm{E}(Y)$$

Definition 4.3 *(B&E, Def.2.4.1)*

The **variance** (Dutch: variantie), notated by $\mathrm{Var}(X)$ or σ^2 of a continuous random variable X is defined by:

$$\sigma^2 = \mathrm{Var}(X) = \mathrm{E}((X-\mu)^2) = \int_{-\infty}^{\infty} (x-\mu)^2\, f(x)\, dx$$

Theorem 4.3

If X and Y independent random variables and a and b are any constants, then:

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - \mu^2 \quad (= \mathrm{E}(X^2) - (\mathrm{E}(X))^2)$$
$$\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X)$$
$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$$

The k-th **moment** of a random variable X is defined by $\mathrm{E}(X^k)$.

The k-th **central moment** of a random variable X is defined by $\mathrm{E}((X-\mu)^k)$.

Theorem 4.4

If X is a random variable with expectation μ ($< \infty$) and standard deviation σ ($< \infty$), and c and k are constants ($c > 0$, $k > 1$), and $g(\cdot)$ is a real valued function, then:

Markov's inequality: $P(|X| \geq c) \leq \dfrac{\mathrm{E}(|X|)}{c}$

Chebyshev's inequality: $P(|X-\mu| \geq k\sigma) \leq \dfrac{1}{k^2}$

Jensen's inequality: If $g(\cdot)$ is convex: $g(\mathrm{E}(X)) \leq \mathrm{E}(g(X))$

Unexpected

As stated above already, it is possible that the variance or even the expectation of a random variable does not exist. A few examples follow.

Example 4.2

Define X as a random variable with density function

$$f_X(x) = \frac{2}{x^3} \qquad \text{on } [1, \infty)$$

This is a proper pdf, with an surface area under its graph equal to 1:

$$\int_1^\infty \frac{2}{x^3}\,dx = \left[-x^{-2}\right]_1^\infty = 1$$

The expected value exists:

$$\mathrm{E}(X) = \int_1^\infty x\,\frac{2}{x^3}\,dx = \left[-2x^{-1}\right]_1^\infty = 2$$

But when we try to determine the variance, we encounter a problem:

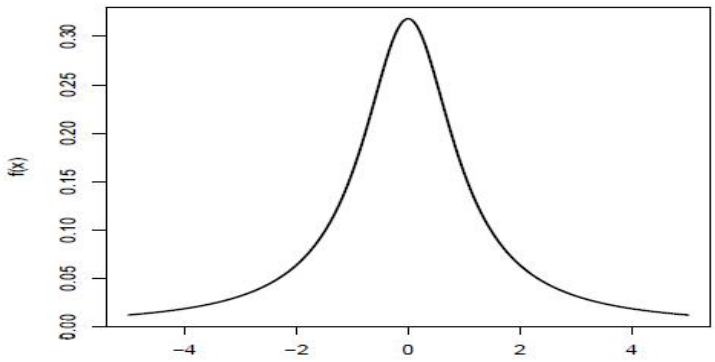
$$\mathrm{E}(X^2) = \int_1^\infty x^2\,\frac{2}{x^3}\,dx = \int_1^\infty \frac{2}{x}\,dx = \left[2\ln x\right]_1^\infty$$

We say that the integral is divergent, and the variance does not exist. One might say as well that the variance is infinitely large. ◄

Example 4.3

The Cauchy distribution is a nice symmetrical unimodal distribution which looks like a normal distribution. It has the following probability density function:

$$f_X(x) = \frac{1}{\pi(x^2+1)} \quad \text{for } x \in (-\infty\,;\,\infty)$$



The surface area under the curvegraph is 1 as it should be:

$$\int_{-\infty}^\infty \frac{1}{\pi(x^2+1)}\,dx = \left[\frac{1}{\pi}\arctan x\right]_{-\infty}^\infty = 1$$

At first sight, most people will probably guess that the mean is equal to 0. However, that is not correct; to see this, we will try to evaluate the integral for the mean:

$$\mathrm{E}(X) = \int_{-\infty}^\infty \frac{x}{\pi(x^2+1)}\,dx = \int_{-\infty}^0 \frac{x}{\pi(x^2+1)}\,dx + \int_0^\infty \frac{x}{\pi(x^2+1)}\,dx = -\infty + \infty = \textit{undefined}$$

(the integral does not converge absolutely). This has the strange consequence that the average of the observations of a random sample will never converge to any value (so also not to 0), even when the sample size becomes larger and larger. In other words, the Law of Large Numbers does not work here.

The Cauchy distribution occurs for example in the following situation. Consider the point $x = 0$ and $y = 1$ in a coordinate system to be the rotation point for an infinitely long straight line. Suppose this line can be given a swing, and it will rotate until it stops at an arbitrary angle. Then the random variable *X,* defined as the spot where this line intersects with the horizontal axis, can be shown to have the Cauchy distribution. *X* can occasionally attain such huge positive or negative values that it becomes impossible to speak about the mean of *X.* (For students interested in physics: this situation is directly related with Huygens' principle in optics, and also has its applications in quantum mechanics).

In the course Probability Theory and Statistics 2, we will show that the ratio of two independent standard normal random variables follows a Cauchy distribution as well. ◄

Symmetry

> ***Definition 4.4*** *(B&E, Def.2.3.5)*
> A probability density function is called **symmetric around *c*** if a constant *c* exists, such that:
> $f(c-x) = f(c+x)$ for all x

In Chapter 1, we discussed the median, percentiles, quantiles and mode. In a straightforward way, we can also use the terms median, percentiles, quantiles and mode of a random variable. For example, the median of X is the solution m of the equation $F(m) = P(X \le m) = 0.5$. The p-th percentile (with p = 1, 2, …, 99) is the value x_p for which $F(x_p) = p/100$.

The mode is the value at which the pdf reaches its maximum.

Example 4.4

Consider the random variable X as defined in Example 4.1. Determine the 90-th percentile. We look for the solution to the third order equation $F(x_{90}) = 0.6x_{90} - 0.1x_{90}^3 + 0.5 = 0.90$. Using a computer or calculator we find: $x_{90} = 0.732$, so: $P(X \le 0.732) = 0.90$. ◄

4.4 Moment generating function

(B&E, Page 78-82)

As we have seen in section 3.5, the probability generating function is a useful tool when dealing with discrete random variables (with the natural numbers as support). Now, we will introduce the moment generating function, which will turn out to be useful for many random variables, both continuous as well as discrete.

> ***Definition 4.5*** *(B&E, Def. 2.5.1)*
> The **moment generating function** (mgf) of a random variable X is defined by:
> $M_X(t) = \mathrm{E}(e^{tX})$ (alternative notation $M[X;t]$)
> (for values of values of t around 0)

Note that for a nonnegative integer-valued random variable X, the mgf is equal to:

$$M_X(t) = \mathrm{E}(e^{tX}) = \Sigma\, e^{tx}\, P(X = x) = \Sigma\, (e^t)^x\, P(X = x) = G_X(e^t)$$

where $G_X(t)$ is the probability generating function of X. This relationship makes it very easy to find the mgf when the probability generating function has already been determined, just by replacing t everywhere by e^t.

Example 4.5

In the previous chapter, the probability generating function for the binomial distribution was found: $G(t) = (pt + q)^n$. This leads directly to the mgf $M(t) = (pe^t + q)^n$. ◄

Example 4.6

Let X be a continuous random variable with $f(x) = e^{-x}$ for $x > 0$. The mgf is:

$$M_X(t) = \int_0^\infty e^{tx} e^{-x}\,dx = \int_0^\infty e^{(t-1)x}\,dx = \left[\frac{1}{t-1} e^{(t-1)x}\right]_0^\infty = \frac{1}{1-t}$$

(The last step is only valid if $t < 1$; for $t \ge 1$ the integral does not converge). ◄

When we replace t in the mgf by the value 0, the mgf will have the value 1, as can be seen as follows:

$$M_X(0) = \mathrm{E}(e^{0X}) = \mathrm{E}(1) = 1.$$

Example 4.6 is an illustration of the fact that the convergence of the integral is not always guaranteed. This is usually no problem at all, as long as $M_X(t)$ does exist for values of t *in an arbitrarily small interval around 0*. But some distributions may show that even for values of t around 0 convergence does not occur. In that case, we could still use a third type of generating function, called the *characteristic function*, defined by:

$$\mathrm{E}(e^{itX}) = \int_{-\infty}^{\infty} e^{itx} f(x)\,dx$$

where i is the imaginary number $i^2 = -1$. The convergence of this integral is never a problem. We will not use this characteristic function in these series of courses, because the mgf is easier to deal with and works sufficiently well to meet our needs.

In the following proofs, we always assume that X is a continuous random variable; for the case X is discrete, the proofs will be similar (just use sums instead of integrals).

Theorem 4.5

If X is a random variable with moment generating function $M_X(t)$, then:

$$M_X'(0) = \mathrm{E}(X)$$

Proof

$$M_X'(t) = \frac{d}{dt} M_X(t) = \frac{d}{dt} \int_{-\infty}^{\infty} e^{tx} f(x)dx = \int_{-\infty}^{\infty} \frac{d}{dt}\left(e^{tx} f(x)\right)dx = \int_{-\infty}^{\infty} xe^{tx} f(x)dx$$

So $M_X'(0) = \int_{-\infty}^{\infty} x f(x)\,dx = \mathrm{E}(X)$

This can be generalised:

Theorem 4.6 *(B&E, Th.2.5.1)*

If X is a random variable with moment generating function $M_X(t)$, then:

$$M_X^{(k)}(0) = \mathrm{E}(X^k) \quad \text{voor } k = 1, 2, 3, \ldots$$

(assuming the k-th moment exists)

Proof

$$M_X^{(k)}(t) = \left(\frac{d}{dt}\right)^{(k)} M_X(t) = \int_{-\infty}^{\infty} \left(\frac{d}{dt}\right)^{(k)} \left(e^{tx} f(x)\right)dx = \int_{-\infty}^{\infty} x^k e^{tx} f(x)dx$$

So $M_X^{(k)}(0) = \int_{-\infty}^{\infty} x^k f(x)dx = \mathrm{E}(X^k)$

Differentiating the mgf produces the moments of X (see Definition 4.5), hence its name. Another way to show this is:

Theorem 4.7 *(B&E, Def.2.5.1)*

If X is a random variable with moment generating function $M_X(t)$, then:

$$M_X(t) = 1 + \sum_{k=1}^{\infty} \frac{E(X^k)}{k!} t^k$$

Proof

(This proof uses an important result from Calculus, i.e. $e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + ...$)

$$M_X(t) = \mathrm{E}(e^{Xt}) = \int_{-\infty}^{\infty} e^{tx} f(x)dx = \int_{-\infty}^{\infty} (1 + tx + \frac{t^2x^2}{2!} + \frac{t^3x^3}{3!} + ...) f(x)dx$$

$$\int_{-\infty}^{\infty} 1 f(x)dx + \int_{-\infty}^{\infty} tx\, f(x)dx + \int_{-\infty}^{\infty} \frac{t^2x^2}{2!} f(x)dx + ...$$

$$= 1 + t\,\mathrm{E}(X) + t^2\,\frac{\mathrm{E}(X^2)}{2!} + ...$$

Check that Theorem 4.6 follows directly from Theorem 4.7 (by taking the derivatives)!

Theorem 4.8 *(B&E, Def.2.5.2)*

If X is a random variable with moment generating function $M_X(t)$, and the random variable Y is defined by $Y = aX + b$, then:

$$M_Y(t) = e^{bt} M_X(at) \quad \text{(or using alternative notation: } M[aX + b; t] = e^{bt} M[X; at]\text{)}$$

Proof

$$M_Y(t) = \mathrm{E}(e^{tY}) = \mathrm{E}(e^{t(aX+b)}) = e^{bt}\mathrm{E}(e^{(ta)X}) = e^{bt} M_X(at)$$

Example 4.7

Define Y by $Y = 1 + 5X$ where X as defined in Example 4.6. The mgf of Y follows easily by applying Theorem 4.8:

$$M_Y(t) = e^t M_X(5t) = \frac{e^t}{1 - 5t}.$$ ◄

The following theorem is important because it is widely used in many future results and proofs.

Theorem 4.9 *(B&E, Def.5.5.1)*

If X and Y are two independent random variables, then:

$$M_{X+Y}(t) = M_X(t) \cdot M_Y(t)$$

Proof will be postponed until the course Probability Theory and Statistics 2.

Moment generating functions have the same pleasant property as probability generating functions: they are unique for each distribution, which means that if two random variables have the same mgf, then those two random variables have the same probability distribution. However, in general it is not a straightforward task to derive the pdf from a specific given mgf. But for many types of frequently encountered distributions, the mgf's are known (see for example the last page of this reader). So if the mgf for a particular random variable can be determined, and if this resulting mgf can then be recognized as the mgf of a known type of probability distribution, then we have actually identified the probability distribution of the random variable. You will see many of those applications of the mgf; below a very simple example:

Example 4.8

If $X \sim \mathrm{BIN}(n, p)$ and $Y \sim \mathrm{BIN}(m, p)$, then the mgf's are (see Example 4.5): $M_X(t) = (pe^t + q)^n$ and $M_Y(t) = (pe^t + q)^m$. What is then the probability distribution of the sum $X + Y$, assuming that X and Y are independent? Using mgf's, it follows from Theorem 4.8 that $M_{X+Y}(t) = (pe^t + q)^n (pe^t + q)^m = (pe^t + q)^{n+m}$. The result can now be recognised as the mgf of a binomial distributed random variable, which shows that $X + Y \sim \mathrm{BIN}(n + m,\ p)$. ◄

4.5 Special continuous distributions

(B&E, Page 109-124, 268-276)

4.5.1 The uniform distribution

The uniform or homogeneous distribution has a constant density (over a specific interval), so that the graph of the pdf looks like a rectangle. This distribution is the continuous analogue to the discrete uniform distribution, and is applicable in situations where there is no preference for any value within a given interval. If the support of X is the interval (a, b) (whether the boundaries a and b are included or not is of no relevance), the density at this interval is equal to $(b-a)^{-1}$, which ensures that the surface area under the pdf equals 1.

Example 4.9

Someone is waiting at a subway station for the next train. The metro train has doors at a constant distance of 8 metres apart (from midpoint to midpoint). If a train stops at the station, what is then the distribution of the distance of the waiting passenger to the midpoint of the nearest door? It is clear that the distance will never exceed four metres, but any value within the range of 0 to 4 metres is just as likely as any other value. So $f(x) = 1/4 \quad \text{for } x \in [0,\ 4]$. ◄

Definition 4.6 *(B&E, Page 109)*

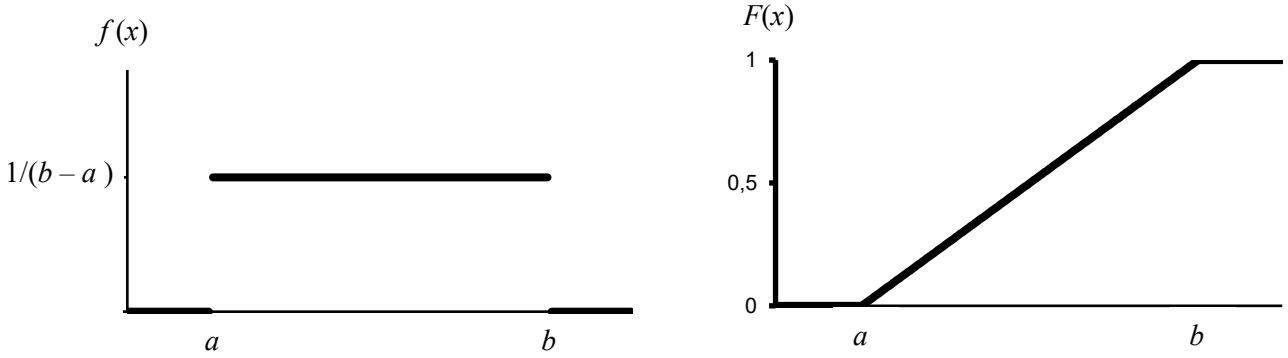
A continuous random variable X has a **(continuous) uniform distribution** on the interval (a, b) (notation $X \sim \text{UNIF}(a,b)$) if its probability density function is given by:

$$f(x) = \frac{1}{b-a} \qquad \text{for } a < x < b$$

The CDF can then be found to be (check!):

$$F(x) = \begin{cases} 0 & \text{for } x \le a \\ \dfrac{x-a}{b-a} & \text{for } x \in (a\ ,\ b) \\ 1 & \text{for } x \ge b \end{cases}$$

In can easily be seen that the CDF is indeed continuous as it should be, by looking specifically at the points a and b (no jumps should occur there). We can draw the following figures:



Moment generating function

$$M_X(t) = \text{E}(e^{tX}) = \int_{-\infty}^{\infty} e^{tx} f(x)\, dx = \int_a^b e^{tx} \frac{1}{b-a}\, dx = \frac{1}{b-a}\left(e^{tx} \Big/ t \right)\Big|_{x=a}^{x=b} = \frac{e^{tb} - e^{ta}}{t(b-a)}$$

Expected value

Using the symmetry of the distribution, it is simple to see that the expectation should be exactly at the midpoint of the interval (a, b), so

$$\mathrm{E}(X) = \tfrac{1}{2}(a + b)$$

Of course, we could have found this answer by integration:

$$\mathrm{E}(X) = \int_a^b x\,(\frac{1}{b-a})\,dx = \left[\frac{x^2}{2(b-a)}\right]_a^b = \frac{b^2-a^2}{2(b-a)} = \frac{a+b}{2}$$

Using the derivative of the mgf would have given the same answer, but in this particular case that would have lead to more complicated calculations.

Variance

$$\mathrm{E}(X^2) = \int_a^b x^2\,(\frac{1}{b-a})\,dx = \left[\frac{x^3}{3(b-a)}\right]_a^b = \frac{b^3-a^3}{3(b-a)} = \frac{b^2+ba+a^2}{3}$$

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - \mu^2 = \frac{b^2+ba+a^2}{3} - \frac{(a+b)^2}{4} = \frac{(b-a)^2}{12}$$

The uniform distribution plays a major role in simulation, where the computer must draw (or simulate) random numbers from a specific probability distribution. Simulation software always include a so-called random number generator. Such a random number generator always generates numbers from a uniform distribution on the interval (0, 1). In section 4.6.4 , we will discuss how those uniformly distributed numbers can be transformed into numbers that can be interpreted as being drawn from any desired distribution.

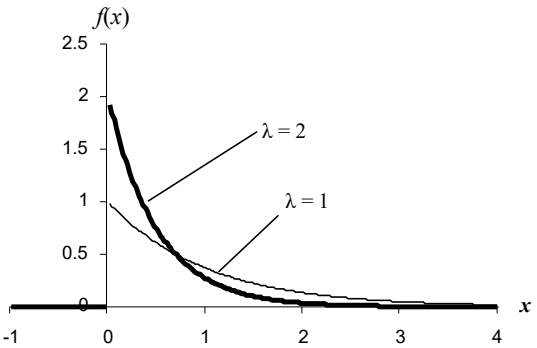
4.5.2 The exponential distribution

Definition 4.7 *(B&E, Page 115)*

A continuous random variable X has an **exponential distribution** with parameter λ (notation $X \sim \mathrm{EXP}(\lambda)$) if its probability density function is given by:

$$f(x) = \lambda e^{-\lambda x} \qquad \text{for } x > 0$$

where $\lambda > 0$.



CDF

$$F(x) = \int_0^x \lambda e^{-\lambda u}\,du = -e^{-\lambda u}\Big|_0^x = 1 - e^{-\lambda x} \qquad \text{for } x \geq 0$$

Moment generating function

$$M_X(\mathrm{t}) = \mathrm{E}(e^{tX}) = \int_0^\infty e^{tx}\ \lambda e^{-\lambda x}\ dx = \lambda \int_0^\infty e^{-(\lambda - t)x}\ dx$$

$$= \frac{\lambda}{\lambda - t} \int_0^\infty (\lambda - t) e^{-(\lambda - t)x}\ dx = \frac{\lambda}{\lambda - t}$$

(The integral in the last line is equal to 1, since it is the area under the curve of a pdf for an exponential distribution with parameter $\lambda - t$, but only if $\lambda - t > 0$, or $t < \lambda$. Because $\lambda > 0$, we can see that the mgf exists for values of t around 0).

Expectation

Using partial integration, we find:

$$\mathrm{E}(X) = \int_0^\infty x\ (\lambda e^{-\lambda x})\ dx = \int_0^\infty -x\ d(e^{-\lambda x})$$

$$= \left[-x\ e^{-\lambda x}\right]_0^\infty - \int_0^\infty -e^{-\lambda x}\ d(x) = 0 + \left[-\frac{1}{\lambda} e^{-\lambda x}\right]_0^\infty = \frac{1}{\lambda}$$

Or using the mgf:

$$M'_X(t) = \lambda(\lambda - t)^{-2} \quad \Rightarrow \quad \mathrm{E}(X) = M'_X(0) = \lambda(\lambda)^{-2} = \frac{1}{\lambda}$$

Variance

We first determine $\mathrm{E}(X^2)$ by partial integration (or use the mgf):

$$\mathrm{E}(X^2) = \int_0^\infty x^2\ (\lambda e^{-\lambda x})\ dx = \int_0^\infty -x^2\ d(e^{-\lambda x})$$

$$= \left[-x^2\ e^{-\lambda x}\right]_0^\infty - \int_0^\infty -e^{-\lambda x}\ d(x^2) = 0 + \int_0^\infty 2xe^{-\lambda x}\ dx \overset{p.i.}{=} \frac{2}{\lambda^2}$$

And thus:

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - \mu^2 = \frac{2}{\lambda^2} - (\frac{1}{\lambda})^2 = \frac{1}{\lambda^2}$$

Remark: Definition 4.7 gives the formula for the exponential distribution as is most commonly used in textbooks. From the formulas above, it can be seen easily that the smaller λ, the greater the expectation and variance. But by replacing the intensity parameter λ by 1/θ, as is done in Bain & Engelhardt, the following pdf results: $f(x) = 1/\theta\ e^{-x/\theta}$ with notation $X \sim \mathrm{EXP}(\theta)$. The parameter θ represents then the expectation (as well as the standard deviation). Note that a distribution which we denote by EXP(2), is notated by Bain & Engelhardt by EXP(1/2)! Some computer packages use the first definition, and others the second, so make sure that you are aware of these differences.

Derivation of the formula for the density function

The exponential distribution is frequently used as a model for waiting times or lifespans. This can be explained by the direct relationship between the exponential distribution and the Poisson process (see Section 3.7.7), which we will now show. We know that in a Poisson process with intensity λ, the number of events in a time interval of length t is Poisson distributed with parameter λt. Now we will focus on the time X that passes between two *consecutive* events in a Poisson process. Saying that the time X is longer than x time units is the very same thing as saying that no events took place within an interval of length x. If we define Y as the number of events in a period with length x, then we know that $Y \sim \mathrm{POI}(\lambda x)$, and we can write:

$$P(X > x) = P(Y = 0 \mid Y \sim \mathrm{POI}(\lambda x)) = e^{-\lambda x} \frac{(\lambda x)^0}{0!} = e^{-\lambda x}.$$

From this result, we can derive the CDF and the pdf of X:

$$F_X(x) = P(X \le x) = 1 - P(X > x) = 1 - e^{-\lambda x}$$
$$f_X(x) = F_X'(x) = \lambda e^{-\lambda x}$$

This is the probability density function of the exponential distribution; so the time between two consecutive events in a Poisson process (with intensity parameter λ) is exponentially distributed (with the same parameter λ)!

Because of the memoryless property (which will be proven directly below), the distribution of the remaining waiting time will be the same *irrespective* of how much time has already passed since the last event in a Poisson process. So the time between two consecutive events has the same distribution as the distribution of time until the next event happens if we start observing a Poisson process at any arbitrary moment. Therefore, we can also say that the distribution of time until the first event happens in a Poisson process is exponentially distributed.

Memoryless property

In section 3.7.5 we showed that the (discrete) geometric distribution has the memoryless property. The exponential distribution is the only continuous distribution with this property.

Theorem 4.10 (memorylessness of the exponential distribution) *(B&E, Th. 3.3.3)*

If $X \sim \text{EXP}(\lambda)$, then (for $a, b > 0$):

$$P(X > a + b \mid X > a) = P(X > b)$$

Proof

$$P(X > a + b \mid X > a) = \frac{P(X > a + b \cap X > a)}{P(X > a)}$$
$$= \frac{P(X > a + b)}{P(X > a)} = \frac{e^{-\lambda(a+b)}}{e^{-\lambda a}} = e^{-\lambda b} = P(X > b)$$

Example 4.10

A part in an electronic device has an exponentially distributed service life with expected value of two years. Suppose that the component already functions well for 3 years, what is the probability that the component has a *total* life span of more than four years?

Answer: $P(X > 4 \mid X > 3) = P(X > 1) = e^{-1/2} = 0.606$. ◄

Example 4.11 Failure rate of a system

The reliability $R(t)$ of a system is often defined as the probability that the service life, often notated by T, will exceed the value t.

$$R(t) = P(T > t)$$

Note that

$$R(t) = P(T > t) = 1 - F(t)$$

In practice, one is often interested in the question whether a system that has already survived for t time units, is about to fail. Or, in other words, the density function of the service life distribution, given that the system has been working up to that time. This can be expressed as the conditional density function of T given $T > t$ and is called the *failure intensity* or *hazard rate*:

$$h(t) = \frac{f(t)}{R(t)} = \frac{f(t)}{1 - F(t)}$$

A system with an exponential lifespan distribution has a constant failure rate, because we get:

$$h(t) = \frac{f(t)}{R(t)} = \frac{f(t)}{1-F(t)} = \frac{\lambda e^{-\lambda t}}{1-(1-e^{-\lambda t})} = \frac{\lambda e^{-\lambda t}}{e^{-\lambda t}} = \lambda$$

This is of course closely related to the memoryless property: it does not matter for the remaining life span to know how long the system already functions properly. ◄

The two-parameter exponential distribution

It is sometimes convenient to shift the pdf of an exponential distribution, for example when X represents a waiting time which is always at least η minutes:

Definition 4.8 *(B&E, Ex. 3.4.1)*

A continuous random variable X has a **two-parameter exponential distribution** with parameters λ and η (notation $X \sim \text{EXP}(\lambda, \eta)$) if its probability density function is given by:

$$f(x) = \lambda e^{-\lambda(x-\eta)} \qquad \text{for } x > \eta$$

where $\lambda > 0$.

Check for yourself that: $M_X(t) = \dfrac{\lambda e^{\eta t}}{\lambda - t}$, $\text{E}(X) = \eta + \dfrac{1}{\lambda}$ and $\text{Var}(X) = \dfrac{1}{\lambda^2}$.

4.5.3 The gamma distribution

Again we will consider a Poisson process with intensity λ; as we have seen in the previous section, the waiting time until the first event is exponentially distributed with parameter λ. Now we will derive the pdf of X, where X is defined as the waiting time until the r-th event happens. First, we will determine the probability $P(X > x)$. If $X > x$, then the r-th event has not yet occurred at time x. Therefore, in the interval $(0, x)$ at most $r-1$ events can have occurred. The number of events (notated by Y) in the interval $(0, x)$ has a Poisson distribution with expectation λx, so:

$$P(X > x) = P(Y \le r-1 \mid Y \sim \text{POI}(\lambda x)) = \sum_{k=0}^{r-1} e^{-\lambda x} \frac{(\lambda x)^k}{k!}$$

Note: X is continuous (waiting *time* until the r-th event), but Y is discrete (*number* of events). Differentiating $F(x) = 1 - P(X > x)$ gives the pdf of X:

$$f(x) = -\frac{d}{dx}\left(\sum_{k=0}^{r-1} e^{-\lambda x} \frac{(\lambda x)^k}{k!}\right) = -\frac{d}{dx}\left(e^{-\lambda x} + \sum_{k=1}^{r-1} e^{-\lambda x} \frac{(\lambda x)^k}{k!}\right) =$$

$$= \lambda e^{-\lambda x} + \sum_{k=1}^{r-1}\left(\lambda e^{-\lambda x} \frac{(\lambda x)^k}{k!} - e^{-\lambda x} \frac{k(\lambda x)^{k-1}\lambda}{k!}\right) =$$

$$= \lambda e^{-\lambda x} + \lambda e^{-\lambda x} \sum_{k=1}^{r-1}\left(\frac{(\lambda x)^k}{k!} - \frac{(\lambda x)^{k-1}}{(k-1)!}\right) = \lambda e^{-\lambda x} \frac{(\lambda x)^{r-1}}{(r-1)!} = \frac{\lambda^r}{(r-1)!} x^{r-1} e^{-\lambda x}$$

Check for yourself that the first step in the third line follows directly because almost all terms cancel each other out.

In the following definition we replace λ by $1/\theta$, which then results in the most common definition for the gamma distribution (used by Bain & Engelhardt as well). Moreover, we generalise this distribution to allow also non-integer values for r:

Definition 4.9 *(B&E, Page 111)*

A continuous random variable X has a **gamma distribution** with parameters θ and r (notation $X \sim \text{GAM}(\theta, r)$) if its probability density function is given by:

$$f(x) = \frac{1}{\theta^r \Gamma(r)} x^{r-1} e^{-x/\theta} \qquad \text{for } x > 0$$

where $\theta > 0$ and $r > 0$.

Note that GAM(θ, r =1) and EXP(1/θ) denote the very same distribution.

The formula above for the pdf shows in the denominator the so-called gamma function (hence the name of this distribution; see also Appendix A). This function is defined by:

$$\Gamma(r) = \int_0^\infty x^{r-1} e^{-x} dx \quad (r > 0)$$

Whenever r is an integer, we can use partial integration to show that $\Gamma(r) = (r-1)!$ (so $(r-1)$-factorial). We can view the gamma-function as well as the necessary scaling factor which makes sure that the surface area under the pdf of the gamma distribution is equal to 1:

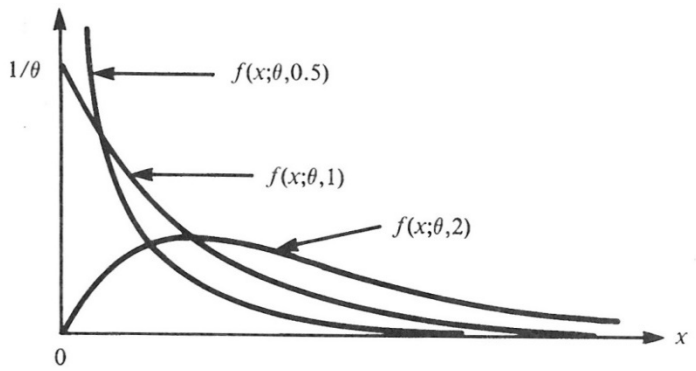
$$\int_0^\infty f(x)dx = 1$$

$$\Rightarrow \quad \frac{1}{\theta^r \Gamma(r)} \int_0^\infty x^{r-1} e^{-x/\theta} dx = 1 \qquad \text{(for the next step, substitute } x \text{ by } u\theta)$$

$$\Rightarrow \quad \frac{1}{\theta^r \Gamma(r)} \int_0^\infty (u\theta)^{r-1} e^{-u} \theta du = 1$$

$$\Rightarrow \quad \frac{1}{\Gamma(r)} \int_0^\infty u^{r-1} e^{-u} du = 1 \quad \Rightarrow \quad \Gamma(r) = \int_0^\infty u^{r-1} e^{-u} du$$

The figure below shows the pdf's for three different values of r; as can be seen, it determines the shape of the distribution, and is therefore called a shape parameter.



The parameter θ is called a scale parameter; the shape itself remains the same, but it determines how much the distribution is 'stretched' (see also the figure for two pdf's for the exponential distribution on page 86).

Moment generating function

The mgf can be found by integration using the pdf (see exercises). Here, we will use another way for the special case that r is integer-valued. The waiting time X until the r-th event in a Poisson process can then be interpreted as the sum of the r waiting times between consecutive events, so

$X = W_1 + \ldots + W_r$, where each W_i is exponentially distributed. The mgf for each W_i is (see previous section): $M_{W_i}(t) = \dfrac{\lambda}{\lambda - t}$. Furthermore, all those r waiting times will be independent random variables, so we can apply Theorem 4.9 repeatedly:

$$M_X(t) = M_{W_1+\cdots+W_r}(t) \;=\; M_{W_1}(t)\cdot M_{W_2}(t)\cdots M_{W_r}(t) \;=\; \left(\frac{\lambda}{\lambda - t}\right)^r \quad (\text{if } t < \lambda)$$

After replacing λ by $1/\theta$, we get: $M_X(t) = \left(\dfrac{1/\theta}{1/\theta - t}\right)^r = \left(\dfrac{1}{1-\theta t}\right)^r$.

In exercise 4.46 you will be asked to prove this result as well for the case that r is not an integer.

Expected value and variance

These can simply be derived from the mgf (check for yourself!):

$$\mathrm{E}(X) = r / \lambda = r\theta$$
$$\mathrm{Var}(X) = r / \lambda^2 = r\theta^2$$

For integer values for r, these results also follow by writing X again as the sum of exponentially distributed random variables and applying Theorem 4.2 and Theorem 4.3.

4.5.4 The normal distribution

In the 17-th century when Galileo Galilei tried to measure the distances to stars, he noticed that the size of measurement errors was distributed in a particular way: small errors were more likely than large errors, and positive errors were just as likely as negative errors. In the 18-th century many biologists registered the weight, height and other characteristics of many individuals of some species and very often the same distribution showed up. De Moivre, Laplace, Adrain and Gauss have made great contributions to the 'discovery' of the so-called normal distribution. But only in 1915 Fisher gave the formula for the probability density function in the form that we use today.

It also became clear that the normal distribution also shows up as the limiting distribution in many situations. For example, the binomial, the Poisson and hypergeometric distributions all start to look similar to the normal distribution under certain conditions.

It can be demonstrated that the normal distribution occurs if a random variable can be viewed as being the result of many small and (almost) independent factors. For example, that is indeed the case for the height of the male (or female) population, since the height is the resultant of all kinds of genetic influences, eating patterns in the youth, diseases etc. In the course Probability Theory and Statistics 3 it will also be proved that the sample mean, provided that n is sufficiently large, is approximately normally distributed, even when the population itself is not normally distributed at all (the very important Central Limit Theorem).

Definition 4.10 *(B&E, Page 118)*

A continuous random variable X has a **normal distribution** with parameters μ and σ (notation $X \sim \mathrm{N}(\mu, \sigma^2)$) if its probability density function is given by:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \qquad \text{for } x \in (-\infty\,;\,\infty)$$

where $\sigma > 0$.

Remark: Some authors and computer packages use the somewhat different notation $X \sim \mathrm{N}(\mu\,,\,\sigma)$)!

Note that the parameters are notated by the symbols μ and σ^2, which seems to suggest that these parameters are equal to the expected value and the variance respectively. That is indeed the case, as will be shown a little later in this section.

Without much effort we can see that this pdf has its maximum at $x = \mu$ and that the density is symmetrical with respect to $x = \mu$. The CDF

$$F(x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{u-\mu}{\sigma})^2} du \qquad \text{for } x \in (-\infty, \infty)$$

cannot be given in the form of an analytic function; it can only be evaluated by numerical approximation. In the tables you will find only the CDF of the so-called **standard normal** distribution, which is a normal distribution with $\mu = 0$ and $\sigma^2 = 1$. We use very frequently the symbol Z to display a standard normal distribution, and the Greek character phi to indicate the corresponding pdf and CDF: $Z \sim N(0, 1)$ with pdf

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \qquad \text{for } z \in (-\infty, \infty)$$

and CDF

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} du \qquad \text{for } z \in (-\infty\,;\,\infty)$$

Using a table to find cumulative probabilities for the standard normal distribution

Until around 25 years ago, the use of a table was the only way to find probabilities for normally distributed random variables, because - as said - the integral for the CDF could only be evaluated numerically, and that had been done so very conscientiously by the constructors of those tables. Today, of course, it is much easier to use a computer program (for example Excel; but even Excel uses very detailed tables to find those values). However, a student in probability theory and statistics should still learn how to use such old-fashioned tables. Partly in order to find answers during an examination, but more importantly, because the use of those tables improves the understanding of the subject matter. Below is a shortened table in which many lines are omitted; in the appendix of this reader you can find the complete table.

Table Cumulative probabilities for the standard normal distribution

z	0.00	0.01	0.02	0.03	**0.04**	0.05	0.06	0.07	0.08	0.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	**.6331**	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
...										
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
...										
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
...										
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
...										
3.7	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999

Say we need to find the probability $\Phi(0.34) = P(Z \le 0.34)$. At $z = 0.34$ (= 0.3 + 0.04) the value for the CDF can be found directly in the table above in row 0.3 and column 0.04: 0.6331. This probability is sometimes called the 'left tail probability' (Dutch: *linkeroverschrijdingskans* van 0.34).

If we need to find a probability that Z is *greater than (or equal to)* a particular value (the right tail probability; Dutch: rechteroverschrijdingskans), like for example $P(Z \ge 0.41)$, then we use the complement rule: $P(Z \ge 0.41) = 1 - P(Z \le 0.41) = 1 - 0.6591 = 0.3409$.

Note that the table only shows probabilities for positive z values. That is sufficient, because the standard normal distribution is symmetrical around 0. For example: $P(Z \le -0.41) = P(Z \ge 0.41) = 0.3409$. And $P(Z \ge -1.07) = P(Z \le 1.07) = 0.8577$.

Often, we need to solve an inverse problem, like finding the value z_α for which $P(Z \ge z_\alpha) = \alpha$, with α a specific given value between 0 and 1. Note that z_α as defined here equals the (1– α)-quantile of the standard normal distribution. For example, for $\alpha = 0.01$ we see that $z_{0.01}$ is equal to the 0.99 quantile, since $P(Z \ge z_{0.01}) = 0.01$ implies that $P(Z \le z_{0.01}) = 0.99$. With the table above, we can determine that $z_{0.01}$ must be a value somewhere between 2.32 and 2.33 (check!). At the bottom of the standard normal table as shown at the end of this reader, a small second table gives the values of z_α for some frequently encountered values of α (like 0.01, 0.025, 0.05 and 0.10). Because we will often need values like z_α when we perform a statistical hypothesis test (see course Probability Theory and Statistics 2) in order to identify a so-called critical region, these values z_α are often called (right tail) critical values.

Remark: Bain & Engelhardt defines z_α as the α-quantile of the standard normal distribution, so $P(Z \le z_\alpha) = \alpha$ (instead of the (1–α)-quantile as we do here). Textbooks are divided in this respect, so the reader is strongly advised to be aware of these differences! The definition we use here leads to a more simple notation for confidence intervals and critical regions, see next course.

The next, very important, theorem shows that any normally distributed random variable can be transformed to a standard normally distributed random variable, and vice-versa.

Theorem 4.11 *(≈B&E, Th. 3.3.4)*

If $X \sim N(\mu, \sigma^2)$ and Z is defined by $Z = \frac{X - \mu}{\sigma}$, then $Z \sim N(0, 1)$.

And if $Z \sim N(0, 1)$ and X is defined by $X = \sigma Z + \mu$, then $X \sim N(\mu, \sigma^2)$.

Proof

For the first part of the proof, we determine the CDF of Z:

$$F_Z(z) = P(Z \le z) = P(\frac{X - \mu}{\sigma} \le z) = P(X \le \mu + z\sigma) = F_X(\mu + z\sigma)$$

We can then determine the pdf by taking the derivative of the CDF:

$$f_Z(z) = \frac{dF_Z(z)}{dz} = \frac{dF_X(\mu + z\sigma)}{dz} = f_X(\mu + z\sigma) \cdot \sigma \quad \text{(using the chain rule)}$$

Substituting $(\mu + z\sigma)$ into the pdf of X results in:

$$f_Z(z) = f_X(\mu + z\sigma) \cdot \sigma = \frac{\sigma}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{\mu + z\sigma - \mu}{\sigma})^2} = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$$

Because this is indeed the pdf for the standard normal distribution, this completes the proof. The proof for the second part is very similar (do it yourself!)

The first transformation in Theorem 4.11, so from any normal distribution to a standard normal distribution, is called **standardisation**. By standardising, we only need a table for the standard normal distribution to find (within a certain accuracy) any probability from any normal distribution.

Example 4.12

A random variable X has a normal distribution with expectation $\mu = 500$ and standard deviation $\sigma = 5$. Determine the probability that X will be greater than 501.7.

$$P(X > 501.7) = P(\frac{X - \mu}{\sigma} > \frac{501.7 - 500}{5})$$

$$= P(Z > 0.34) = 1 - \Phi(0.34) = 1 - 0.6331 = 0.3669$$ ◄

Moment generating function

For the *standard normal* distribution, the mgf can be found as follows:

$$M_Z(t) = \mathrm{E}(e^{tZ}) = \int_{-\infty}^{\infty} e^{tz} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}\, dz = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}[z^2-2tz]}\, dz$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}[(z-t)^2-t^2]}\, dz$$

$$= e^{\frac{1}{2}t^2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z-t)^2}\, dz = e^{\frac{1}{2}t^2}$$

The integral in the last line is simply the surface area under the pdf for a normal distribution with expectation t and variance 1, which **must** be equal to 1 for any value of t.

To find the mgf for $X \sim \mathrm{N}(\mu, \sigma^2)$, we first use Theorem 4.11 to write X as $\sigma Z + \mu$ (with $Z \sim \mathrm{N}(0, 1)$) , followed by applying Theorem 4.8 and using the mgf for Z:

$$M_X(t) = M_{\sigma Z+\mu}(t) = e^{\mu t} M_Z(\sigma t) = e^{\mu t} e^{\sigma^2 t^2/2} = e^{\mu t + \sigma^2 t^2/2}.$$

Below, we will use the mgf to prove a generalised version of Theorem 4.11:

Theorem 4.12

If $X \sim \mathrm{N}(\mu, \sigma^2)$, and $Y = aX + b$, then $Y \sim \mathrm{N}(a\mu+b, a^2\sigma^2)$.

Proof

The mgf for Y can be found using Theorem 4.8:

$$M_Y(t) = e^{bt} M_X(at) = e^{bt} e^{\mu at + a^2\sigma^2 t^2/2} = e^{(\mu a+b)t + a^2\sigma^2 t^2/2}$$

This result can be recognised as the mgf of a $\mathrm{N}(a\mu+b, a^2\sigma^2)$ distributed random variable.

Expectation

It seems obvious that the point of symmetry (μ) of the pdf of $X \sim \mathrm{N}(\mu, \sigma^2)$ is indeed the expected value of X. But can we also verify this formally? We will start with the integral in Definition 4.10 to find the expected value for the *standard normal* distribution:

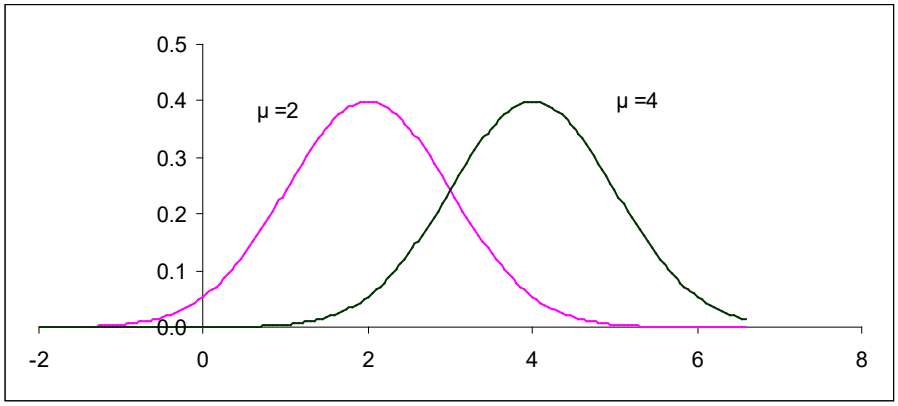
$$\mathrm{E}(Z) = \int_{-\infty}^{\infty} \frac{z}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}\, dz = \left[\frac{-1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \right]_{-\infty}^{\infty} = 0 - 0 = 0$$

If $X \sim \mathrm{N}(\mu, \sigma^2)$, then X can be written as $X = \sigma Z + \mu$. Theorem 4.2 gives then the result:
$\mathrm{E}(X) = \mathrm{E}(\sigma Z + \mu) = \sigma\, \mathrm{E}(Z) + \mu = \mu$, exactly what we 'expected'.

Even simpler is to determine the expectation with the help of the mgf:

$$M_X'(t) = e^{\mu t + \sigma^2 t^2/2}(\mu + \sigma^2 t) \quad \Rightarrow \quad \mathrm{E}(X) = M_X'(0) = \mu$$

The figure below shows two normal pdf's, both with $\sigma = 1$, which clearly illustrates why μ is also called a location parameter.



Variance

The determination of the variance using integrals is a bit more complicated. We start with $\mathrm{E}(Z^2)$:

$$\mathrm{E}(Z^2) = \int_{-\infty}^{\infty} \frac{z^2}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}\,dz = \int_{-\infty}^{\infty} \frac{-z}{\sqrt{2\pi}}\,d\left(e^{-\frac{1}{2}z^2}\right)$$

$$= \left[\frac{-z}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}\right]_{-\infty}^{\infty} - \int_{-\infty}^{\infty} e^{-\frac{1}{2}z^2}\,d\left(\frac{-z}{\sqrt{2\pi}}\right) =$$

$$= 0 + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}z^2}\,dz = 1 \text{ (because it is the surface area under the pdf of } Z\text{).}$$

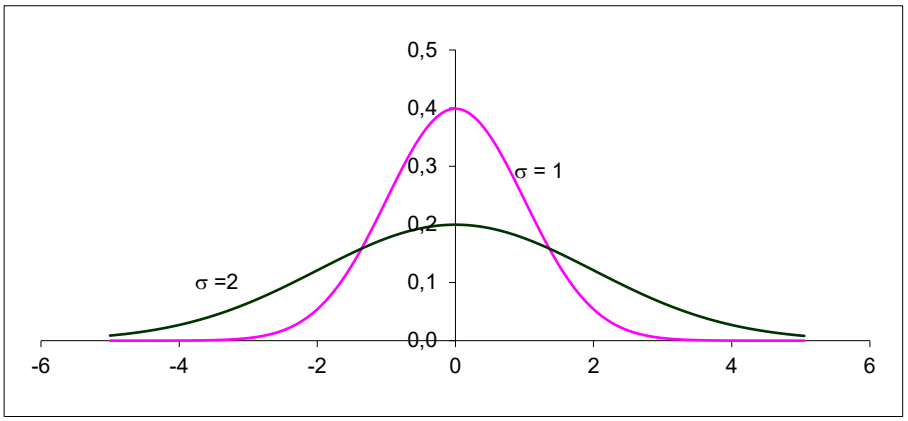
We could use the above when determining the variance of X:

$$\mathrm{Var}(X) = \mathrm{E}(X-\mu)^2 = \int_{-\infty}^{\infty} \frac{(x-\mu)^2}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}\,dx \overset{\text{subst } z=\frac{x-\mu}{\sigma}}{=} \sigma^2 \int_{-\infty}^{\infty} \frac{z^2}{\sqrt{2\pi}} e^{-\frac{1}{2}(z)^2}\,dz = \sigma^2$$

(Or, alternatively, we can first show that Var(Z) = 1, followed by applying Theorem 4.3 to the equation $X = \sigma Z + \mu$).

In exercise 4.59, you will be asked to use the mgf for the derivation of the variance.

The figure below illustrates why σ can be called a scale parameter (two distributions, both with μ = 0).



Sums of normally distributed random variables

Theorem 4.13 *(≈B&E, Ex. 6.4.7)*

If X and Y are two independent random variables, with $X \sim \mathrm{N}(\mu_X, \sigma_X^2)$ and $Y \sim \mathrm{N}(\mu_Y, \sigma_Y^2)$, then the sum of X and Y is normally distributed as: $X + Y \sim \mathrm{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$.

Proof

We have seen already that $M_X(t) = e^{\mu_X t + \sigma_X^2 t^2/2}$ and $M_Y(t) = e^{\mu_Y t + \sigma_Y^2 t^2/2}$. Application of Theorem 4.9 gives:

$$M_{X+Y}(t) = e^{\mu_X t + \sigma_X^2 t^2/2} \cdot e^{\mu_Y t + \sigma_Y^2 t^2/2} = e^{(\mu_X + \mu_Y)t + (\sigma_X^2 + \sigma_Y^2)t^2/2}$$

This expression can easily be recognised as the mgf of a normal distribution with expectation equal to the sum of the expectations of X and Y, and variance equal to the sum of the variances of X and Y.

In the course Probability Theory and Statistics 2 we will discuss the distributions of sample statistics (like the sample mean) when a sample is taken from a specific distribution. We can give an important

result here already: when applying Theorem 4.13 repeatedly to the sum of n independent normally distributed random variables $X_1,...,X_n$, each with the same expectation μ and variance σ^2, we get:

$$\sum_{i=1}^{n} X_i \sim \mathrm{N}(n\mu,\ n\sigma^2)$$

By dividing this sum by n, the sample mean is obtained. This sample mean thus has a normal distribution with the same expectation as the distribution from which the observations are derived (i.e. μ), and with variance σ^2/n (see Theorem 4.12 with $a = 1/n$ and $b = 0$), so:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim \mathrm{N}(\mu,\ \sigma^2/n)$$

Note that the variance of the sample mean decreases when the sample size increases, which means that the sample mean will be more and more concentrated around the population mean μ for increasing values of n.

The *Central-Limit Theorem* (for which the proof will have to wait until the course Probability Theory and Statistics 3) says that this distribution will be also the distribution - but only by approximation - of the mean of a row of random variables with expectation μ and standard deviation σ, *even if the population from which the sample was drawn is not normally distributed*! What 'by approximation' implies, is not to say in a few words. When the population distribution is symmetric, then the approximation is already quite accurate for the sum of a relatively small number of random variables (e.g., $n = 10$). Often the rule of thumb is used that the Central Limit Theorem gives good approximations whenever $n \geq 30$.

Normal approximations for discrete distributions

The normal approximation to the binomial distribution

The binomial distribution arises when we look at the number of successes in a series of n independent trials of an experiment, each trial with the same probability p of success (Bernoulli trials). If X_i is the outcome of the i-th trial, the outcome of the sequence $X_1,\ X_2, \ldots, X_n$ will be a row of 0's and 1's. The sum X $(= X_1 + X_2 + \ldots + X_n)$ of this sequence is equal to the number of 1's and is, in other words, equal to the number of successes in the sequence of Bernoulli trials, and thus $X \sim$ BIN (n, p). So we can apply the Central Limit Theorem to show that X will be approximately normally distributed, as long as n is large enough. This can also be seen when looking at histograms of binomial distributions, which start to look quite similar to the form of the pdf of a normal distribution (at least when n is large enough and p is not too small or too large). As a rule one can say that the approximation is 'fairly good' if both $np \geq 5$ and $nq \geq 5$.

We will perform such an approximation as follows: say $X \sim$ BIN(n, p). We will approximate X by a normally distributed random variable X^c with *the same expectation* $(=np)$ *and variance* $(=npq)$ as X, so $X^c \sim \mathrm{N}(np,\ npq)$ (check!). If n sufficiently large, then:

$$P(X \leq x) \approx P(X^c \leq x).$$

We can evaluate this probability by standardising:

$$P(X \leq x) \approx P(X^c \leq x) \ = \ P\left(\frac{X^c - np}{\sqrt{npq}} \leq \frac{x - np}{\sqrt{npq}}\right) \ = \ P\left(Z \leq \frac{x - np}{\sqrt{npq}}\right)$$

For any x, n and p, we can use the standard normal distribution table to find these probabilities.

Continuity correction

The quality of the approximation can be improved by using a so-called continuity correction. Because X can only attain integer values, the following three expressions are equivalent when x is an integer:

$$P(X \leq x) \ = \ P(X < x + \tfrac{1}{2}) = \ P(X < x + 1)$$

However, if we would approximate these three probabilities by the method above, we would get three different approximations. Usually, the middle one will give the best results. Therefore, we will get:

$$P(X \le x) \approx P(Z \le \frac{x + \frac{1}{2} - np}{\sqrt{npq}})$$

Here, we applied the continuity correction by adding ½ before standardising the normal probability. Using a similar argument, we can see that the best approximation for $P(X \ge x)$ with x an integer is:

$$P(X \ge x) \approx P(Z \ge \frac{x - \frac{1}{2} - np}{\sqrt{npq}})$$

Example 4.13

Of a certain kind of tulip bulbs 90% will produce flowering plants (independent of each other, and also regardless of the weather). Calculate the probability with the aid of the normal approximation that 20 bulbs will produce at most 17 flowering plants. Define X as the number of flowering plants, so $X \sim$ BIN (20, 0.9). The expected value for X is 20 × 0.9 = 18, and the variance is 20 × 0.9 × 0.1.

$$P(X \le 17) = P(X \le 17\tfrac{1}{2})$$

$$\approx P\left(Z \le \frac{17½ - 18}{\sqrt{20(0.9)(0.1)}}\right) = P\left(Z \le \frac{-0.5}{1.3416}\right) = \qquad \text{(with } Z \sim \text{N(0, 1))}$$

$$= P\left(Z \le -0.372\ldots\right) \approx P\left(Z \le -0.37\right) = \ 0.3557$$

The exact binomial probability is 0.3231 (check). This is a case where the normal approximation (0.3557) is not very close to the correct value (0.3231). Partly, this is due to the rounding to two decimals of 0.372 to 0.37 in the last line of the approximation. But much more importantly: the rule of thumb that both $np \ge 5$ and $nq \ge 5$ is not satisfied. Whenever the probability of success is closer to ½, the approximation will be better.

We now calculate the probability of at least 11 flowering plants when the probability of success is 0.6:

$$P(X \ge 11) = P(X \le 10\tfrac{1}{2})$$

$$\approx P\left(Z \ge \frac{10½ - 12}{\sqrt{20(0.6)(0.4)}}\right) = P\left(Z \ge \frac{-1.5}{2.1908}\right) =$$

$$\approx P\left(Z \ \ge -0.68\right) \ = \ 0.7517$$

The exact binomial probability is now 0.7553, and we can see that this probability is quite close to the exact value now the rule of thumb is satisfied. ◄

In general: the larger the value of n and the closer the value of p to ½, the better the approximation. The continuity correction (±½) almost always improves the quality of the approximation.

Individual probabilities can be found as follows:

$$P\left(X = x\right) = P(x - ½ < X < x + ½) = P\left(X < x + ½\right) - P(X < x - ½) \approx P\left(X^c \le x + ½\right) - P(X^c \le x - ½)$$

Example 4.14

$P(X = 10 \mid X \sim \text{BIN}(20, ½)) = P(9½ < X < 10½) = P(X < 10½) - P(X < 9½) =$

$$\approx P\left(Z \le \frac{10½ - 10}{\sqrt{20(½)(½)}}\right) - P\left(Z \le \frac{9½ - 10}{\sqrt{20(½)(½)}}\right) = P\left(Z \le \frac{0.5}{2.236}\right) - P\left(Z \le \frac{-0.5}{2.236}\right) =$$

$$= P(Z \le 0.2236) - P(Z \le -0.2236) \approx P(Z \le 0.22) - P(Z \le -0.22) = 0.1742$$

The exact binomial probability is 0.1762. ◄

The normal approximation to the Poisson distribution

Probabilities from the Poisson distribution can be approximated by normal probabilities in an entirely analogous manner as above. A Poisson distribution looks more and more symmetrical when the value of its parameter μ increases. Because the normal distribution is always symmetrical, this already tells us that the value of μ should not be too small for a reasonably good quality approximation. As a rule of thumb, many authors state that the value of μ should be at least 15.

Say $X \sim$ POI(μ). Again, we will approximate X by a normally distributed random variable X^c which has the same expected value and variance as X, so $X^c \sim \mathrm{N}(\mu,\ \mu)$. Using the continuity correction, we get:

$$P(X \le x) \approx P(X^c \le x + ½) \ = \ P(\frac{X^c - \mu}{\sqrt{\mu}} \le \frac{x + ½ - \mu}{\sqrt{\mu}}) \ = \ P(Z \le \frac{x + ½ - \mu}{\sqrt{\mu}})$$

Example 4.15

Let X have a Poisson distribution with μ = 15 (the last value which usually is still shown in tables). Assume that we need to find the probability $P(X \ge 14)$:

$$P(X \ge 14) \approx P(Z \ge \frac{13 + ½ - 15}{\sqrt{15}})$$

$$= P\left(Z \ge -0.3873\right) \approx 1 - P\left(Z < -0.39\right) = \ 0.6517$$

This approximation is reasonably close to the exact value of 0.6368.
As an example for finding a probability of one specific outcome:

$$P\left(X = 13\right) = P\left(X < 13.5\right) - P\left(X < 12.5\right)$$

$$\approx P(Z \le \frac{13.5 - 15}{\sqrt{15}}) - P(Z \le \frac{12.5 - 15}{\sqrt{15}})$$

$$= P(Z < -0.3873) - P(Z < -0.6455)$$

$$\approx 0.3483 - 0.2578 = 0.0905$$

(compared to the exact Poisson probability of 0.0956) ◄

4.5.5 The chi-square distribution

The chi-square distribution is a special case of the gamma distribution: if X has a gamma distribution with θ = 2 and r = ν/2, then we can also say that X has a chi-square distribution with parameter ν. The pdf can be derived directly from the pdf of the gamma distribution:

Definition 4.11

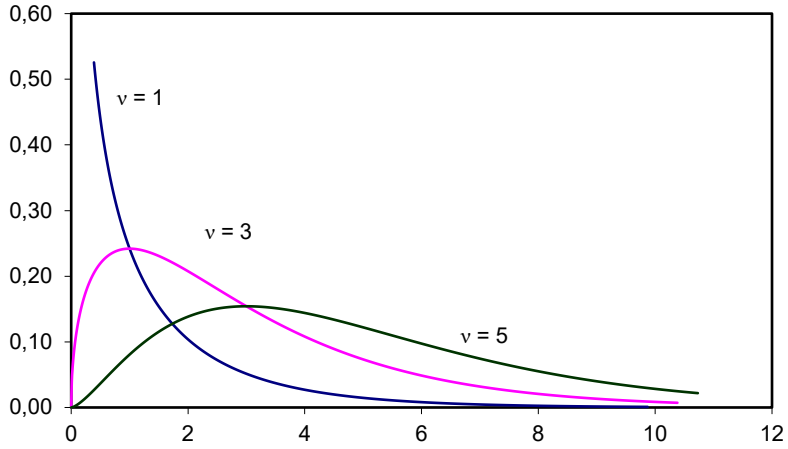
A continuous random variable X has a **chi-square distribution** with parameter ν (notation $X \sim \chi^2(\nu)$) if its probability density function is given by:

$$f(x) = \frac{e^{-\frac{1}{2}x}\, x^{\frac{1}{2}\nu - 1}}{2^{\frac{1}{2}\nu}\, \Gamma(\frac{1}{2}\nu)} \qquad \text{for } x > 0$$

where ν = 1, 2, 3, … .

Note that the Greek letter chi χ is used here, hence the name for this distribution. For the parameter of the chi-square distribution, often the Greek letter ν ('nu') is used, which is usually called the degrees of freedom. The reason for that name will only become clear during the next course. Please check for yourself that the chi-square distribution with two degrees of freedom is the same as the exponential distribution with expected value 2!

The figure below shows the curves for the pdf's of χ^2-distributions with $\nu = 1$, 3 and 5. The first curve never touches the vertical axis, the second has its top (mode) at 1, the third at 3 (= $\nu - 2$ according to exercise 4.78).



The reason that the chi-square distribution is discussed on its own (and not just as a special case of the gamma distribution) is that it will be encountered very frequently. That is mainly due to the following theorem:

Theorem 4.14 *(B&E, Th. 8.3.5)*

If Z is a standard normal random variable, then $X = Z^2$ has a chi-square distribution with 1 degree of freedom.

Proof

We will first determine the CDF of X (note that X can never attain negative values, so $x \geq 0$):

$$F_X(x) = P(X \leq x) = P(Z^2 \leq x)$$

$$= P(-\sqrt{x} \leq Z \leq \sqrt{x}) = \Phi(\sqrt{x}) - \Phi(-\sqrt{x}) = 2\Phi(\sqrt{x}) - 1$$

By differentiating (chain rule!) the pdf is obtained:

$$f_X(x) = 2\phi(\sqrt{x})\left(\frac{d\sqrt{x}}{dx}\right) = 2\phi(\sqrt{x})\tfrac{1}{2}x^{-\frac{1}{2}} =$$

$$= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(\sqrt{x})^2} x^{-\frac{1}{2}} = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x} x^{-\frac{1}{2}}$$

By comparing this result with the formula in Definition 4.11, we can see that this is actually the pdf of a chi-squared distributed random variable with 1 degree of freedom (recall that $\Gamma(\frac{1}{2}) = \sqrt{\pi}$, see also Appendix A.4).

An alternative proof can be given by using mgf's (see later in Example Example 4.23).

In exercise 4.74 you will be asked to show, if X are Y independent and $X \sim \chi^2(m)$ and $Y \sim \chi^2(n)$, that the sum of X and Y is also chi-square distributed: $X + Y \sim \chi^2(m+n)$.

The direct consequence is that the sum of squares of m independent standard normally distributed random variables follows a chi-square distribution with m degrees of freedom:

$$Z_1^2 + Z_2^2 + Z_3^2 + \ldots + Z_m^2 \sim \chi^2(m)$$

Moment generating function, expectation and variance

The mgf, expectation and variance for the chi-square distribution can be obtained directly from those of the gamma distribution. If $X \sim \chi^2(\nu)$:

$$M_X(t) = \frac{1}{(1-2t)^{\nu/2}} = (1-2t)^{-\nu/2}$$

$$\mathrm{E}(X) = \nu$$

$$\mathrm{Var}(X) = 2\nu$$

Table

A table for the chi-square distribution usually consists of one page with critical values. Each line in the table refers to a number of degrees of freedom. The table entries for the table in the Appendix (see page 125) show right-tail critical values (i.e. $\chi^2_\alpha(\nu)$ such $P(\chi^2(\nu) > \chi^2_\alpha(\nu)) = \alpha$ for a number of prefixed values for α. For example, we find in the third row and in the column for α = 0.10 that $P(\chi^2 > 6.251 \mid \nu = 3) = 0.100$.

Remark 1. Some tables (for example those in Bain & Engelhardt) are based on the left-tail critical values (or simply called the percentiles); in that case we could have found the same value 6.251 by looking in the column for γ = 0.90 such that $P(\chi^2 < 6.251 \mid \nu = 3) = 0.90$.

Remark 2. When you look at the figure above for ν = 3, it might seem that the surface area to the left of 6.251 is less than 10% of the overall area under the pdf; however, that tail of the chi-square distribution is relatively 'thick'.

Using a simple linear transformation, any gamma-distributed random variable can be transformed into a chi-square distributed random variable:

Theorem 4.15 *(B&E, Th. 8.3.3)*

If $X \sim$ GAM(θ, r) and $2r$ is a positive integer, then $Y = 2X/\theta$ has a chi-square distribution with $2r$ degrees of freedom.

Proof

We supply a proof here based on mgf's. We know that (see section 4.5.3): $M_X(t) = (1-\theta t)^{-r}$.

The mgf of Y can be found by applying Theorem 4.8:

$$M_Y(t) = M_X(2t/\theta) = (1-2t)^{-r} = (1-2t)^{-2r/2}$$

This mgf can indeed be recognised as the mgf for a chi-square distribution with $2r$ degrees of freedom.

Example 4.16

Say $X \sim$ GAM(θ=12, r =7), and we are asked to find the 90-th percentile of the distribution of X, i.e. the value $p_{0.90}$ such that $P(X \le p_{0.90}) = 0.90$. So:

$$P(X \le p_{0.90}) = P(2X/12 \le 2p_{0.90}/12) = P(\chi^2(14) \le p_{0.90}/6) = 0.90$$

$$\Rightarrow \quad P(\chi^2(14) > p_{0.90}/6) = 0.10$$, and we find in the table that $P(\chi^2(14) > 21.064) = 0.10$.

Solving $p_{0.90}/6 = 21.064$ gives $p_{0.90} = 126.384$ ◄

4.5.6 The Weibull distribution

The Weibull distribution is widely used to model service lives or lifespans of materials or instruments. It is a kind of generalisation of the exponential distribution. The CDF of the last distribution is:

$$P(X \le x) = F(x) = 1 - e^{-\lambda x} .$$

If we provide the exponent of a power then we arrive at the Weibull-distribution: $F(x) = 1 - e^{-(\lambda x)^\beta}$.

Usually, we replace the inverse of the rate parameter λ by the scale parameter θ (= 1/λ), then:

$F(x) = 1 - e^{-(x/\theta)^\beta}$. By differentiating we get the pdf:

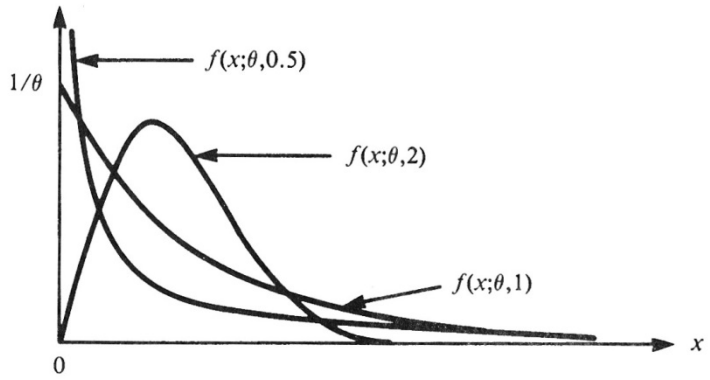
Definition 4.12

A continuous random variable X has a **Weibull-distribution** with parameters θ and β (notation $X \sim \text{WEI}(\theta, \beta)$) if its probability density function is given by:

$$f(x) = \frac{\beta\, x^{\beta-1}}{\theta^{\beta}} e^{-(\frac{x}{\theta})^{\beta}} \qquad \text{for } x > 0$$

where $\theta > 0$ and $\beta > 0$.

Note that a Weibull distribution with $\beta = 1$ is the same as an exponential distribution.



Expectation and variance

Without proof we give the following results here. Interested people can find the derivation at page 117 of Bain&Engelhardt.

$$\text{E}(X) = \theta\, \Gamma\left(\frac{1}{\beta}+1\right)$$

$$\text{Var}(X) = \theta^2 \left(\Gamma\left(\frac{1}{\beta}+1\right) - \Gamma\left(\frac{1}{\beta}+1\right)^2 \right)$$

The moment generating function can be derived, but due to its very complicated structure of not much use in practical applications.

4.6 Transformation (= function) of a random variable

(B&E, Page 194-204)

Introduction

Often, we will have to deal with distributions of transformations of random variables. As an illustration, let us focus on a simple transformation of a random variable with a simple distribution. Let X be uniformly distributed on the interval [0 , 2], so the density of X is constant over this interval. A random sample of five observations from this distribution could possibly result in:

0.13 1.41 1.56 1.12 0.39

Consider the transformation $Y = X^2$. Observations on Y can be obtained by squaring the observations on X. The corresponding row of observations for Y would then be:

0.0169 1.9881 2.4336 1.2544 0.1521

The support of Y is clearly the interval [0, 4]; only on this interval, Y will have a positive density. It is easy to see that as a result of squaring the values of X, small outcomes of X (<1) will be 'pushed' towards 0, while larger outcomes of X (>1) will be stretched out along the interval from 1 to 4. Intuitively, this means that the density of Y for small values of y will be higher than for larger values of y. See Example 4.17 for a mathematical derivation of the pdf of Y.

In general, the problem we are facing is how to determine the pdf of a random variable $Y = u(X)$ where $u(\cdot)$ is a real-valued function and the pdf $f_X(x)$ of X is known. Three different methods will be discussed here which can be helpful in the derivation of the pdf of Y. In the next section, we will discuss the so-called CDF method. The second method is the *transformation* method, and the third uses moment generating functions.

4.6.1 CDF-method

This method carries the name CDF method, because it focuses on trying to express the CDF of Y in terms of the CDF of X. If we are successful and Y is continuous, then we should take the derivative of the CDF in order to obtain the pdf of Y. In general, it is a good idea to start with the determination of the support of Y. Of course, only the behaviour of the function $u(x)$ for values of x belonging to the support of X needs to be taken into account. We will define S_X as the support set of X, so $S_X = \{x \mid f_X(x) > 0\}$, and S_Y as the support set of Y, so $S_Y = \{y \mid f_Y(y) > 0\}$. S_Y is logically the same as the image of S_X under the function $u(x)$ (or $S_Y = \{y \mid y = u(x)$ for some $x \in S_X\}$), and is often very simple to find.

Example 4.17

Suppose $Y = X^2$ and $X \sim$ UNIF(0,2). Because X can only have values between 0 and 2, it is immediately obvious that $Y = X^2$ can only attain values on the interval (0,4). Now, we will try to write the CDF of Y in such a way that we arrive at an expression using the CDF of X:

$$F_Y(y) = P(Y \le y) = P(X^2 \le y) = P(-\sqrt{y} \le X \le \sqrt{y}) = P(X \le \sqrt{y}) = F_X(\sqrt{y})$$

(Note that at the fourth step above, we use the fact that X can never attain negative values).
The CDF of X follows from the fact that $X \sim$ UNIF(0, 2), so: $F_X(x) = \frac{1}{2}x$. Together with the above result, we obtain:

$$F_Y(y) = F_X(\sqrt{y}) = \tfrac{1}{2}\sqrt{y}$$

The density of Y follows by taking the derivative of the CDF:

$$f_Y(y) = \frac{dF_Y(y)}{dy} = \frac{d}{dy}\left(\tfrac{1}{2}\sqrt{y}\right) = \frac{1}{4\sqrt{y}} \qquad \text{for } 0 < y < 4$$

We see that the density is huge for values of Y very close to 0, while it is around 1/8 for values of Y close to 4, which is indeed according to the intuitive idea discussed in the introduction above. ◄

See the proofs of Theorem 4.11 and Theorem 4.14 for two other examples of the CDF-method.

A few examples of functions of X:
$Y = 2X$: $F_Y(y) = P(Y \le y) = P(2X \le y) = P(X \le y/2) = F_X(y/2)$
$Y = \sqrt{X}$: $F_Y(y) = P(Y \le y) = P(\sqrt{X} \le y) = P(X \le y^2) = F_X(y^2)$ (only if $P(X < 0) = 0$)
$Y = |X|$: $F_Y(y) = P(Y \le y) = P(|X| \le y) = P(-y \le X \le y) = F_X(y) - F_X(-y)$

For further evaluation of the probabilities above, it is important to know how the support of X is mapped onto the support of Y. For situations where this mapping is not immediately clear, drawing a graph of the function $y = u(x)$ (for $x \in S_X$) can be very helpful. The support of Y can then be found simply by looking at the range of values for y. But the graph will also show whether two different values of x can be mapped onto the same value of y. In the latter case, $y = u(x)$ is *not* a one-to-one function (on S_X to S_Y). We will say that $y = u(x)$ is a one-to-one function if the equation $y = u(x)$ has a unique solution $x = u^{-1}(y)$. We need to be extra careful in situations where the transformation is not one-to-one, as can be seen in the next example.

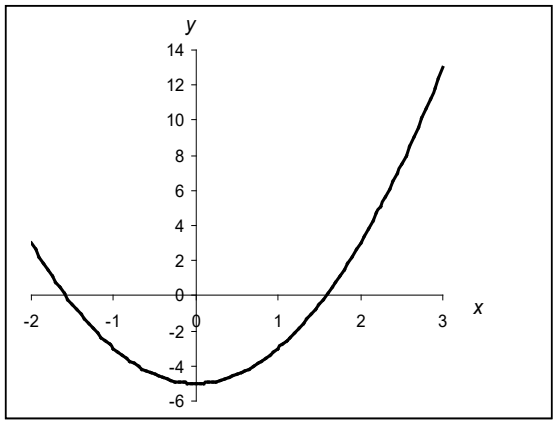
Example 4.18

Let X have a uniform distribution on the interval [-2, 3]. We will determine the pdf of Y, with Y defined by $Y = 2X^2 - 5$. We could start with the CDF of Y:

$$F_Y(y) = P(Y \le y) = P(2X^2 - 5 \le y) = P(X^2 \le \frac{y+5}{2})$$

$$= P(-\sqrt{\frac{y+5}{2}} \le X \le +\sqrt{\frac{y+5}{2}}) \;=\; F_X(\sqrt{\frac{y+5}{2}}) - F_X(-\sqrt{\frac{y+5}{2}})$$

But when we try to evaluate these last probabilities, we have to be very much aware of the support sets. Drawing the curve for the function $y = 2x^2 - 5$ for all values of $x \in [-2,3]$ makes it more clear.



It is now immediately clear that the support of Y is the interval from -5 (when x=0) to 13 (when x=3). So $S_Y = \{y \mid -5 \le y \le 13\}$, which is also the projection of the curve on the vertical axis. The graph also shows that the function $y = 2x^2 - 5$ is not one-to-one, since for each y between −5 and 3, two different values for x map onto the same y value.

Now for $-5 < y < 3$, we still have: $F_Y(y) = F_X(\sqrt{\frac{y+5}{2}}) - F_X(-\sqrt{\frac{y+5}{2}})$, but for $3 < y < 13$, we get: $F_Y(y) = F_X(\sqrt{\frac{y+5}{2}}) - F_X(-\sqrt{\frac{y+5}{2}}) = F_X(\sqrt{\frac{y+5}{2}})$ (because the second part is 0).

Because X ~UNIF[-2, 3], we know that $F_X(x) = \frac{1}{5}(x+2)$ (for $-2 \le x \le 3$), and we obtain:

For $-5 < y < 3$: $F_Y(y) = F_X(\sqrt{\frac{y+5}{2}}) - F_X(-\sqrt{\frac{y+5}{2}}) = = \frac{2}{5}\sqrt{\frac{y+5}{2}}$

Differentiating results in the pdf: $f_Y(y) = \frac{1}{5\sqrt{2(y+5)}}$ for $y \in (-5\ ,\ 3)$

For $3 < y < 13$: $F_Y(y) = F_X(\sqrt{\frac{y+5}{2}}) \;=\; \frac{1}{5}(\sqrt{\frac{y+5}{2}} + 2)$

which leads to the pdf: $f_Y(y) = \frac{1}{10\sqrt{2(y+5)}}$ for $y \in (3\ ,\ 13)$

All together: $f_Y(y) = \begin{cases} \frac{1}{5\sqrt{2(y+5)}} & \text{for } -5 < y \le 3 \\ \frac{1}{10\sqrt{2(y+5)}} & \text{for } 3 < y \le 13 \end{cases}$

◄

4.6.2 Transformation method

The CDF-method discussed in the previous section is mainly convenient for *continuous* random variables. When dealing with *discrete* random variables, the transformation method discussed below can be used more conveniently and we will see that the pdf of $Y = u(X)$ can usually be expressed directly as a function of the pdf of X itself. But also for continuous random variables, we will show here that it is in general not necessary to determine explicitly the CDF of Y first; the transformation method for continuous random variables can be shown to follow directly from the CDF method.

One-to-one transformations

Theorem 4.16 (discrete situation) *(B&E, Th. 6.3.1)*

If X is a discrete random variable with pdf $f_X(x)$, and $Y = u(X)$ defines a one-to-one transformation of $S_X = \{x \mid f_X(x) > 0\}$ to $S_Y = \{y \mid f_Y(y) > 0\}$, then

$$f_Y(y) = f_X(u^{-1}(y)) \qquad \text{for } y \in S_Y$$

Proof

$$f_Y(y) = P(Y = y) = P(u(X) = y) = P(X = u^{-1}(y)) = f_X(u^{-1}(y))$$

Example 4.19

Say $X \sim$ DU(0, 4), so $f_X(x) = 1/5$ for $x = \{0, 1, 2, 3, 4\}$. Define $Y = 2X + 3$. The possible outcomes for Y are 3, 5, 7, 9 and 11, each with probability 1/5, because (for example)

$$f_Y(5) = P(Y = 5) = P(u(X) = 5) = P(2X + 3 = 5) = P(X = 1) = f_X(1) = 1/5$$ ◄

The notation above might seem unnecessarily complex and trivial at the same time. However, that is not so in the situation we are dealing with continuous random variables:

Theorem 4.17 (continuous situation) *(B&E, Th. 6.3.2)*

Let X be a continuous random variable with pdf $f_X(x)$, and $Y = u(X)$ defines a one-to-one transformation of $S_X = \{x \mid f_X(x) > 0\}$ to $S_Y = \{y \mid f_Y(y) > 0\}$. If the derivative $\frac{d}{dy}u^{-1}(y)$ is continuous and nonzero on $y \in S_Y$, then

$$f_Y(y) = f_X(u^{-1}(y))\left|\frac{d}{dy}u^{-1}(y)\right| \qquad \text{for } y \in S_Y$$

Proof

If $y = u(x)$ is a one-to-one function, then $u(x)$ is either a monotonic increasing or a monotonic decreasing function. Assuming first that $u(x)$ is increasing, then $u^{-1}(y)$ must be increasing as well (which in turn means that its derivative is positive) and we obtain:

$$F_Y(y) = P(Y \le y) = P(u(X) \le y) = P(X \le u^{-1}(y)) = F_X(u^{-1}(y))$$

By differentiating with respect to y, it follows (use the chain rule):

$$f_Y(y) = f_X(u^{-1}(y))\frac{d}{dy}u^{-1}(y) = f_X(u^{-1}(y))\left|\frac{d}{dy}u^{-1}(y)\right|$$

(because the derivative of $u^{-1}(y)$ is positive, taking the absolute value is irrelevant.)

If we assume $u(x)$ to be decreasing, then $u^{-1}(y)$ is decreasing as well, and

$$F_Y(y) = P(Y \le y) = P(u(X) \le y) = P(X \ge u^{-1}(y)) = 1 - F_X(u^{-1}(y))$$

$$\Rightarrow \quad f_Y(y) = -f_X(u^{-1}(y))\frac{d}{dy}u^{-1}(y) = f_X(u^{-1}(y))\left|\frac{d}{dy}u^{-1}(y)\right|$$

(note that the derivative of $u^{-1}(y)$ is now negative, which means that taking the absolute value cancels out the minus sign in front of the expression).

Example 4.20

Suppose $X \sim$ UNIF(0, 1), so $f_X(x) = 1$ for $0 < x < 1$. Consider the transformation $Y = -(\ln X)/\lambda$, with $\lambda > 0$. The support set of Y is clearly $\{y \mid y > 0\}$, and the transformation is one-to-one. The inverse of the function $y = u(x) = -(\ln x)/\lambda$ is $x = u^{-1}(y) = e^{-\lambda y}$ (check!). Using Theorem 4.17 we obtain the pdf of Y:

$$f_Y(y) = f_X(u^{-1}(y))\left|\frac{d}{dy}u^{-1}(y)\right| = 1 \cdot \left|-\lambda e^{-\lambda y}\right| = \lambda e^{-\lambda y} \qquad \text{for } y > 0$$

Since this is the pdf for an exponential distribution, it follows that $Y \sim$ EXP(λ). ◀

The most common type of transformation is the linear transformation, $Y = aX + b$. We have already seen that $\mathrm{E}(Y) = a\mathrm{E}(X) + b$, and $\mathrm{Var}(Y) = a^2\mathrm{Var}(X)$. We can also use the transformation method to find a general expression for the pdf of Y. Note first that the function $y = ax + b$ is one-to-one, with the inverse function $x = \dfrac{y-b}{a}$.

When X is a continuous random variable, Theorem 4.17 gives us the result:

$$f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right) \quad \text{for } y \in S_Y = \{y \mid (y-b)/a \in S_X\}$$

And when X is a discrete random variable, Theorem 4.16 gives us the result.

Transformations which are not one-to-one

If $u(x)$ is not one-to-one on $S_X = \{x \mid f_X(x) > 0\}$, then there will not exist a unique solution to the equation $y = u(x)$ for every $x \in S_X$, but, in general, it will be possible to partition S_X into a collection of disjoint subsets $S_{X,i}$ (i = 1, 2,…) such that on each subset the function $y = u(x)$ will be one-to-one, so we can write $x = u_i^{-1}(y)$ as the inverse function of $y = u(x)$ for $x \in S_{X,i}$. The image of each subset $S_{X,i}$ under the function $u(x)$ will be denoted as $S_{Y,i}$.

In the **discrete** situation, we get:

$f_Y(y) = \sum_i f_X(u_i^{-1}(y))$ where the sum is taken over all i such that $y \in S_{Y,i}$

Example 4.21

Assume X takes one of the values 0, 1, 2, 3, 4 and 5 with probabilities $f_X(0)$, $f_X(1)$, $f_X(2)$, $f_X(3)$, $f_X(4)$ and $f_X(5)$ respectively. If $Y = u(X) = (X-2)^2$, then Y can assume the values 0, 1, 4 and 9. The transformation is not one-to-one, but we can split $S_X = \{0, 1, 2, 3, 4, 5\}$ into the two subsets $S_{X,1} = \{0, 1, 2\}$ and $S_{X,2} = \{3, 4, 5\}$, which are mapped onto $S_{Y,1} = \{0, 1, 4\}$ and $S_{Y,2} = \{1, 4, 9\}$. The corresponding inverse functions are: $u_1^{-1}(y) = 2 - \sqrt{y}$ (for $y \in S_{Y,1}$) and

$u_2^{-1}(y) = 2 + \sqrt{y}$ (for $y \in S_{Y,2}$). Thus, we find for example for $y = 4$:

$$f_Y(4) = P(Y = 4) = P((X-2)^2 = 4)$$
$$= P(X = u_1^{-1}(4)) + P(X = u_2^{-1}(4))$$
$$= P(X = 0) + P(X = 4) = f_X(0) + f_X(4)$$ ◀

In the **continuous** case, the formula becomes:

$$f_Y(y) = \sum_i f_X(u_i^{-1}(y)) \left| \frac{d}{dy} u_i^{-1}(y) \right|$$ where the sum is taken over all i such that $y \in S_{Y,i}$

Example 4.22

Again we look at the case of Example 4.18, so $X \sim$ UNIF(–2, 3) and $Y = 2X^2 - 5$. The function $y = 2x^2 - 5$ is not one-to-one on the interval [–2, 3]. But we can split this interval (see also the graph in Example 4.18), into $S_{X,1} = [-2, 0)$ and $S_{X,2} = [0, 3]$, and on each of these two intervals the function is one-to-one with $S_{Y,1} = (-5, 3]$ and $S_{Y,2} = [-5, 13]$). The corresponding inverse functions are: $u_1^{-1}(y) = -\sqrt{\frac{y+5}{2}}$ and $u_2^{-1}(y) = \sqrt{\frac{y+5}{2}}$.

For $y \in [-5, 3]$, y is an element of both $S_{Y,1} = (-5, 3]$ and of $S_{Y,2} = [-5, 13]$, so applying the formula $f_Y(y) = \sum_i f_X(u_i^{-1}(y)) \left| \frac{d}{dy} u_i^{-1}(y) \right|$ leads to: (also recall that $f_X(x) = 1/5$ for $x \in [-2, 3]$):

$$f_Y(y) = f_X(u_1^{-1}(y)) \left| \frac{d}{dy} u_1^{-1}(y) \right| + f_X(u_2^{-1}(y)) \left| \frac{d}{dy} u_2^{-1}(y) \right|$$
$$= \frac{1}{5}\frac{1}{2}\frac{1}{\sqrt{2(y+5)}} + \frac{1}{5}\frac{1}{2}\frac{1}{\sqrt{2(y+5)}} = \frac{1}{5}\frac{1}{\sqrt{2(y+5)}}$$

For $y \in [3, 13]$, y is an element only of $S_{Y,2} = [-5, 13]$ and we obtain:

$$f_Y(y) = f_X(u_2^{-1}(y)) \left| \frac{d}{dy} u_2^{-1}(y) \right| = \frac{1}{10}\frac{1}{\sqrt{2(y+5)}}$$

Of course, these results are identical to those found in Example 4.18. ◀

4.6.3 Moment generating function method

We have already seen that the moment generating function of a random variable uniquely determines its probability distribution. So, if we are able to determine the mgf of $Y = u(X)$ and we are also able to recognise the mgf as the mgf of some type of known probability distribution, then we have identified the probability distribution of Y. We have already used this method in the proof of Theorem 4.15. We will now show the use of this method in an alternative proof of Theorem 4.14 .

Example 4.23

Theorem 4.14 states that, if Z is standard normally distributed, then $X = Z^2$ has a chi-square distribution with 1 degree of freedom. Proof:

$$M_X(t) = \mathrm{E}(e^{tX}) = \mathrm{E}(e^{tZ^2}) = \int_{-\infty}^{\infty} e^{tz^2} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2(1-2t)} dz = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2/(1-2t)^{-1}} dz$$

$$= (1-2t)^{-\frac{1}{2}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}(1-2t)^{-\frac{1}{2}}} e^{-\frac{1}{2}z^2/(1-2t)^{-1}} dz$$

$$= (1-2t)^{-\frac{1}{2}}$$

The integral in the third line is equal to 1, since it is the area under the density function of a random variable with a normal distribution with $\mu = 0$ and $\sigma = (1-2t)^{-\frac{1}{2}}$. The result, $M_X(t) = (1-2t)^{-\frac{1}{2}}$, can be recognised (see last page in this reader) as the mgf of a chi-square distribution. ◄

4.6.4 Integral Transformation and simulation

A very special and useful transformation is the transformation of any continuous random variable X by its own CDF:

Theorem 4.18 (Probability Integral Transformation, PIT) *(B&E, Th. 6.3.3)*

If X is a continuous random variable with CDF $F_X(x)$, then $F_X(X)$ has a uniform distribution on the interval (0, 1), so $U = F_X(X) \sim$ UNIF(0, 1).

Proof

Note that $F_X(x) = P(X \le x)$ is by definition a nondecreasing function. We will give a proof here in case $F_X(x)$ is *strictly* increasing (on the support of X) such that the inverse function $x = F_X^{-1}(u)$ exists.

We will supply three different proofs, using the methods in the previous sections.

1. CDF method: The support for U is clearly the interval (0, 1):

$$F_U(u) = P(U \le u) = P(F_X(X) \le u)$$
$$= P(X \le F_X^{-1}(u)) = F_X(F_X^{-1}(u)) = u \qquad \text{for } 0 < u < 1$$

which can easily be recognised as the CDF of a uniform distribution on the interval (0, 1).

2. Transformation method:

$$f_U(u) = f_X(F_X^{-1}(u)) \left| \frac{d}{du} F_X^{-1}(u) \right| = f_X(F_X^{-1}(u)) \frac{d}{du} F_X^{-1}(u)$$

$$= \frac{dF_X(F_X^{-1}(u))}{du} = \frac{du}{du} = 1 \quad \text{for } 0 < u < 1$$

so again it follows that $U \sim$ UNIF(0, 1).

$f_X(F_X^{-1}(u)) = \frac{dF_X(F_X^{-1}(u))}{du}$

$= \frac{du}{du} = 1.$

3. Mgf method:

$$M_U(t) = \mathrm{E}(e^{tU}) = \mathrm{E}(e^{tF_X(X)}) = \int_{-\infty}^{\infty} e^{tF_X(x)} f_X(x) dx = \frac{e^{tF_X(x)}}{t} \Bigg|_{x=-\infty}^{x=\infty} = \frac{e^t - 1}{t}$$

This is the mgf of a uniform distribution on the interval (0, 1).

In case the CDF is not strictly increasing, the theorem remains valid, but we will not provide a proof here.

Remark: This theorem means that any continuous random variable can be transformed into a random variable with a (standard) uniform distribution. This can provide the basis for testing whether a set of observations may be regarded as originating from a specified distribution, using the idea that we would expect observations from a

uniform distribution to be evenly spread over the interval (0, 1), such that for example approximately 10% of all observations should be between 0 and 0.1. Well-known P-P plots and the Kolmogorov-Smirnov tests use this idea.

The following theorem turns everything around:

Theorem 4.19 (Inverse PIT) *(B&E, Th. 6.3.4)*

Let $F(x)$ be an arbitrary CDF. If $U \sim \text{UNIF}(0, 1)$, then $X = F^{-1}(U)$ has a distribution with $F(x)$ as CDF. (In case $F(x)$ is not strictly increasing, we then define $F^{-1}(u)$ as the smallest value of x such that $F(x) \geq u$.)

Proof

$$F_X(x) = P(X \leq x) = P(F^{-1}(U) \leq x) = P(U \leq F(x)) = F(x)$$

where the last step follows directly from the well-known CDF of the uniform distribution ($P(U \leq x) = x$).

Theorem 4.19 is of great practical importance when simulating numbers using a computer. Any software program that does anything with statistics, including Excel, features a random-number generator, which produces random numbers between 0 and 1. Those numbers can therefore be perceived as a random sample from a uniform distribution on the interval (0, 1). Theorem 4.19 provides us with a method to simulate random samples from any desired probability distribution.

Example 4.24

Simulate a drawing from a probability distribution with pdf $f_X(x) = x^2/3$ (for $-1 < x < 2$).

We need first to determine the CDF: $F_X(x) = \int_{-1}^{x} v^2/3 \, dv = v^3/9\Big|_{-1}^{x} = \frac{x^3+1}{9}$ (for $-1 < x < 2$).

Next, we determine the inverse of this CDF:

$$u = \frac{x^3+1}{9} \quad \Rightarrow \quad x = F_X^{-1}(u) = \sqrt[3]{9u-1}$$

Say for example that the random-number generator U produces the number 0.3862. This will result in a value for X equal to $\sqrt[3]{9 \cdot 0.3862 - 1} = 1.3528$. Note that the value 1.3528 is exactly the 0.3862 quantile of the distribution of X, so $F_X(1.3528) = 0.3862$. In this way we can convert each drawn value of U into values which can then be considered to be drawn from the distribution of X. ◄

The above procedure also works for discrete distributions.

Example 4.25

Simulate the number of dots that results when throwing a fair die. The probability function is $f_X(x) = 1/6$ (voor $x = 1, 2, 3, 4, 5, 6$), and the corresponding CDF is:

$$F_X(x) = \begin{cases} 0 & \text{for } x < 1 \\ \text{floor}(x)/6 & \text{for } 1 \leq x < 6 \\ 1 & \text{for } x \geq 6 \end{cases}$$

(the floor function (or Entier) returns the largest integer less than or equal to x).

The function $F_X^{-1}(u)$ as defined in Theorem 4.19 is then: $F_X^{-1}(u) = \begin{cases} 1 & \text{for } 0 < u \leq 1/6 \\ 2 & \text{for } 1/6 < u \leq 2/6 \\ 3 & \text{for } 2/6 < u \leq 3/6 \\ 4 & \text{for } 3/6 < u \leq 4/6 \\ 5 & \text{for } 4/6 < u \leq 5/6 \\ 6 & \text{for } 5/6 < u \leq 1 \end{cases}$

(Check!) ◄

4.6.5 Location, scale and shape parameters

In discussing the various continuous probability distributions in this chapter we have encountered many different parameters. We can distinguish different groups of parameters.

- **Location parameters** (Dutch: locatieparameters), which enable the shifting of a distribution. A larger value will shift a distribution to the right. Examples: the parameter η of the two-parameter exponential distribution and the parameter μ of the normal distribution.
- **Scale parameters** (Dutch:schaalparameters), which primarily change the spread of a distribution. A larger value will 'stretch' the distribution. Examples: the parameter θ of the gamma and the Weibull distribution and parameter σ of the normal distribution.
- **Intensity (or rate) parameters** (Dutch: intensiteitsparameters) are the inverse of a scale parameter in situations where it is useful to talk about a *rate.* Example: the parameter λ of the exponential distribution.
- **Shape parameters** (Dutch: vormparameters), which change the basic shape of a distribution. Examples: the parameter r of the gamma distribution, the parameter β of the Weibull distribution, the parameter ν of the chi-square distribution.

Distributions of the same type, but with different values for their location and/or scale parameters can be transformed into each other. During the discussion of the normal distribution we saw already that any normal distribution can be transformed into a standard normal distribution. However, that is impossible for distributions with different values for a shape parameter (like the parameter r of the gamma distribution).

Below, we write $F(x;\tau)$ for the CDF of X, where τ represents either one or more parameters. This format does not rule out that X might have other parameters which are not explicitly mentioned. In a similar way, we use the notation $f(x;\tau)$ for the pdf.

Definition 4.13 *(B&E, Def. 3.4.1)*

A parameter η is a **location parameter** for the distribution of X if its CDF satisfies the following equation:

$$F(x;\eta) = F(x-\eta;\ 0)$$

This is equivalent with: $f(x;\eta) = f(x-\eta;\ 0)$

Example 4.26

Say $X \sim N(\mu,\sigma^2)$. Here, μ is a location parameter, because the formula for the pdf does not change when we replace μ by 0 and replace x by $(x-\mu)$. In other words: the density of a normally distributed random variable with parameters μ and σ^2 evaluated at the point x (so $f_X(x;\mu,\sigma^2)$) is equal to the density of a normally distributed random variable with parameters **0** and σ^2 evaluated at the point $x-\mu$ (so $f_X(x-\mu;\ 0,\ \sigma^2)$):

$$f(x;\mu) = f(x-\mu;0) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

◄

Definition 4.14 *(B&E, Def. 3.4.2-3)*

A parameter $\theta > 0$ is a **scale parameter** for the distribution of a continuous random variable X if its CDF satisfies the following equation:

$$F(x;\theta) = F\left(\frac{x}{\theta};\ 1\right)$$

Or, equivalently: $f(x;\theta) = \frac{1}{\theta} f\left(\frac{x}{\theta};\ 1\right)$

If X has a location parameter η as well, then we say that θ is the scale parameter of the pair (η, θ) if its CDF satisfies the following equation:

$$F(x;\eta,\theta) = F\left(\frac{x-\eta}{\theta};\ 0,\ 1\right)$$

Or, equivalently: $f(x;\eta,\theta) = \frac{1}{\theta} f\left(\frac{x-\eta}{\theta};\ 0,1\right)$

Check for yourself that the parameter σ of the normal distribution is indeed according to the definition a scale parameter of the location-scale pair (μ,σ). Similarly, the parameter θ of the gamma distribution is a scale parameter.

Remark: the values a and b of the uniform distribution on the interval do not fit the above scheme; however, if we would reparametrize this distribution by defining the parameters $\alpha = a$ and $\beta = (b - a)$, then we can see that α is a location parameter and β the scale parameter of the pair (α, β).

4.7 Problems

4.1 A random variable X has density $k(x^2 - x^3)$ on $[0, 1]$, elsewhere 0.
- a Find k.
- b Calculate $P(X > 0.75)$.
- c Find the distribution function of X.
- d Calculate $P(X > 0.75 \mid X > 0.5)$.

4.2 Let the density of random variable X be given by $f(x) = k \sin(x)$ on $[0 , \pi]$.
- a Show that $k = ½$.
- b Calculate $P(X \le \pi / 3)$
- c Find b such that $P(X \le b) = 1 / 3$.

4.3 Show that $g(y)$ is a probability density function for every $n = 1, 2, 3, \ldots$ and for $r = 1, 2, \ldots, n$, when

$$g(y) = \begin{cases} \binom{n}{r} r\ y^{r-1} (1-y)^{n-r} & \text{for } 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases}$$

4.4 Show that for a continuous distribution with θ as a point of symmetry the following holds:
- a $P(\theta - x < X < \theta + x) = 2\,F(\theta + x) - 1$
- b $E(X) = \theta$ (that is, if this expected value exists)

4.5 Let the density of a random variable X be given by $f(x) = 3\ e^{-3x}$ on $[0 , \infty)$, elsewhere 0.
- a Determine $P(X > ¼)$.
- b Find the probability that $X > ½$ given that $X > ¼$.
- c Calculate the expected value and the variance of X.

4.6 Let the density of a random variable X be given by $f(x) = x$ on [0 , 1) and $f(x) = 2 - x$ on [1 , 2), elsewhere 0.
- a Find E(X) and Var(X).
- b Find $F(x)$ for all values of x.
- c When three independent observations of the random variable X are obtained, then find the probability that at least two of them will be bigger than ½.

4.7 Let X represent the temperature in °C on a future day in June at 11.00 am. The past has taught that this temperature has an expected value of 18.3 and a standard deviation of 4.1. How would the expected value and the standard deviation of the temperature be reported, if the temperature was measured in Fahrenheit?

4.8 Prove that the variance of the sample mean (of a random sample, where all observations are independent of each other) is equal to σ^2/n, where σ^2 is the variance within the population.

4.9 Prove Theorem 4.2 (except the last line).

4.10 Prove Theorem 4.3 (except the last line).

4.11 Prove Theorem 4.4.

4.12 Given is a random variable X, with density cx^{-2} on the interval $[5 ; \infty)$.
- a Find c.
- b Show that the expected value of X does not exist.

4.13 Let the pdf of the random variable X be given by $f(x) = 4x^3$ on [0 , 1], elsewhere 0.
- a Determine the expected value and the variance.
- b Determine $P(\mu - 2\sigma \le X \le \mu + 2\sigma)$ and compare the outcome with the bound obtained from Chebyshev's inequality.

4.14 Assume that the distance between the target and the hit of a shot is a random variable X with density $f(x) = \frac{3}{4}(1 - x^2)$ on [−1 , 1] and elsewhere 0.
- a Calculate E(X) and Var(X).
- b Calculate $P(|X| \le \frac{1}{2})$ and compare the outcome with the bound obtained from Chebyshev's inequality.

4.15 For which t does the moment generating function of a random variable X with density $3x^{-4}$ on (1 , ∞) exist?

4.16 Find the expected value of a random variable X with moment generating function $M_X(t) = (2 - e^t)^{-1}$. Find the probability generating function and find the pdf of the distribution.

4.17 Let the (discrete) pdf of a random variable X be given by: $f(x) = 2^{-x}$ for $x \in \mathbb{N}$
- a Find the probability generating function and use it to find the expected value and variance of X.
- b Find the moment generating function of X.

4.18 For a certain random variable X the moment generating function is $M_X(t) = \frac{1}{2}e^t + \frac{1}{3}e^{2t} + \frac{1}{6}e^{4t}$.
- a Use the moment generating function to find the expected value and variance of X.
- b Find the probability generating function.
- c Find the pdf of the distribution of X.

4.19 Let $M_X(t)$ and $M_Y(t)$ be the moment generating functions of X and Y respectively.
- a Show that $2\,M_X(t)$ cannot be a moment generating function.
- b Can $0.5\,M_X(t) + 0.5\,M_Y(t)$ be a moment generating function? If yes, why, if no, why not?
- c Can $M_X(t) \times M_Y(t)$ be a moment generating function?

4.20 Let the random variable Y be given by the probabilities: $P(Y = y) = \dfrac{\mu^{\frac{y-10}{2}}}{(\frac{y-10}{2})!}e^{-\mu}$ for $y = 10, 12, 14, \ldots$

Find the moment generating function of Y.

4.21 As remarked earlier, when discussing the negative-binomial distribution, two different definitions for this distribution exist. In one definition the quantity (X) represents the number of experiments needed to get the r-th success, and in the other the random variable (Y) represents the number of failures just before getting the r-th success. Explain the relation between the moment generating functions of X and Y.

4.22 Use mgf's to show that the sum of two independent Poisson-distributed quantities X and Y has a Poisson-distribution as well.

4.23 Use mgf's to find out whether, and if so under which conditions, the sum of two independent binomial quantities has a binomial distribution as well.

4.24 The arrival times of customers at a garage are uniformly distributed on the interval between 8.00 am and 9.00 am. When a customer has not arrived yet at 8:30 am, find the probability that this customer will arrive after 8:45 am.

4.25 A traffic light gives red during 60 seconds, then it gives green for 90 seconds, and then orange for 10 seconds. You arrive at a random moment in time. Calculate the probability that you have to wait 30 seconds at most.

4.26 Let a random variable X be uniformly distributed on the interval $(-2\ ,\ 2)$.
a Find $P(1/X < 2)$.
b Find $P(X^{-2} < 2)$.

4.27 Suppose the independent random variables X and Y are uniformly distributed on the intervals $[0,\ 5]$ and $[1\ ,\ 3]$ respectively. Find the moment generating function of $X + Y$.

4.28 Let the random variable X be uniformly distributed on the interval $(-1\ ,\ 1)$. Find $P(¼ < X^2 < ¾)$.

4.29 Determine the probability that at least two out of four UNIF(0, 10)-distributed observations are larger than 7.

4.30 Let X be uniformly distributed on $(0\ ,\ 12)$. Calculate $P(|X - 6| \leq 0.6\sigma\sqrt{3})$ and compare the outcome with the bound obtained from Chebyshev's inequality.

4.31 Find the expected value and variance of the uniform distribution on $[a\ ,\ b\ ]$ using only the fact (already derived) that for a continuous uniform distribution on $[0\ ,\ 1]$ the expected value is ½ and the variance is 1/12.

4.32 Within a random time interval with length t it appears that 1 Poisson event has occurred ($X = 1$). Show that the moment Y at which the event occurred is uniformly distributed on the interval $[0,\ t]$.

4.33 Calculate the median of the exponential distribution.

4.34 Calculate the probability that out of five observations from an exponential distribution exactly two are smaller than the expected value.

4.35 Suppose the independent random variables X and Y are exponentially distributed with expected values 2 and 3 respectively. Find the moment generating function of $X + Y$. Is $X + Y$ also exponentially distributed?

4.36 Let's assume (not conform to reality) that the lifetime of batteries of a certain type has an exponential distribution. Radio A contains two batteries of which at least one must function. The expected lifetime for both these batteries is 200 hours (independent from each other). In radio B there are four batteries of which at least three must function. This type has an expected lifetime of 400 hours. A radio is required that will work for at least 500 hours. Which of the two radio's has the highest probability to meet this requirement?

4.37 The diameter (in μm) of a certain fibre is a random variable X with EXP(λ)-distribution. Fibres with diameter smaller than 1, generate a yield of 3 milli-euro, and a thicker fibre results in a loss of 1 m€. The production process allows the value of λ (in μm^{-1}) to be changed a little bit. Compare $\lambda = ½$ and $\lambda = ¼$ in terms of expected profits.

4.38 An exponential random variable has a right-tail probability of 0.85 at time 200. Calculate at what time the right-tail probability is still 0.95.

4.39 Find the intensity of a Poisson-process, when a probability equal to 0.05 is given for the event that the waiting time between the first and the second event will be longer than two hours.

4.40 A twilight lamp has three light bulbs of type M, of which the lifetime can be considered to be exponential distributed with an expected value of 1000 hours. Calculate the probability that the twilight lamp will give some light for at least 1200 hours. (For the calculation the lifetimes must be regarded as independent.)

4.41 Under usual circumstances a fuse has a hazard rate (see Example 4.11) of 0.05. Calculate the reliability at $t = 125$. (Note that no time unit was given. So if time is measured in *hours*, then the unit of measurement of the hazard rate is *hours* $^{-1}$. In that case one could say the hazard rate is: 0.05 'per hour'.)

4.42 A machine part has a lifetime with density:

$$f(x) = \begin{cases} x & \text{on the interval } [0, 1] \\ 1/x^3 & \text{for } x > 1 \end{cases}$$

a Find the cumulative distribution function of the lifetime.
b Find the hazard rate (see Example 4.11).
c Does the hazard rate reach a maximum at a certain moment? Derive and/or comment.

4.43 Prove that for the two-parameter exponential distribution it holds that: $M_X(t) = \frac{\lambda e^{\eta t}}{\lambda - t}$, $\mathrm{E}(X) = \eta + \frac{1}{\lambda}$ and $\mathrm{Var}(X) = \frac{1}{\lambda^2}$.

4.44 Extra exercise, difficult. We will now investigate the most simple waiting queue system. In this case we assume that clients enter according to a Poisson process with expected value λ, (so the time between two entering clients is exponentially distributed with parameter λ). Furthermore we assume that there is just one server, and that the service time is exponentially distributed with parameter μ. Of course λ must be smaller than μ, because otherwise in the long run more clients would enter than can possibly be served. We denote the state of the system by looking at the number of clients that is present within the system. Thus (*i*) represents the state in which *i* clients are present in the system (including the client being served).

a Why is it not important to know how long the system is already in a certain state (*i*)?

Define p_i as the probability that at a certain moment the system is in state (*i*). It is easy to see that p_i also represents the fraction of time (in the long run) that there are *i* clients in the system.

b How often (per time unit) will the system change from state (i) to state (i+1)?
And how often (per time unit) will the system change from state (i) to (i-1)?
c It is reasonable that in the long run the number of times (per time unit) that the system changes to state (i) must be equal to the number of times that the system leaves state (i). This is also called: rate in equals rate out. Use this information to set up an equation for state (0). Next, do this for state (1), then for state (2).
d Try to express p_i as a function of p_0.
e Now determine p_0, and next p_i.

4.45 Consider a Poisson process where event A occurs on average once every 15 minutes. For the probabilities below one can often use several distributions. Write out these probabilities, if possible, as Poisson probabilities, exponential probabilities, and/or gamma probabilities.

a the probability that one has to wait at least 25 minutes for the first A;
b the probability that in two hours A will occur at most six times;
c the probability that the fourth event will occur within half an hour;
d the probability that after the fifth event one has to wait at least 25 minutes for the sixth one;
e the probability that after the fifth event one has to wait at least 25 minutes for the seventh one;

4.46 Find the moment generating function of the gamma distribution using the density and the definition of the mgf.

4.47 Let $X \sim \mathrm{GAM}(\theta, r)$. Find $\mathrm{E}(X)$ using the pdf of X.

4.48 Prove the second part of Theorem 4.11.

4.49 Assume that 'IKJU'-scores are distributed as $N(100, 10^2)$. Calculate the probability of an IKJU

a above 123.
b between 90 and 110.

4.50 The electrical resistance of certain parts is normally distributed with $\mu = 0.21$ and $\sigma = 0.045$. For a certain purpose these parts must have a resistance of at most 0.2316. What percentage is useful?

4.51 Rivets with a nominal length of ¾ inch in fact have a normally distributed length, while the parameters depend on the supplier.
Supplier I: $\mu_1 = 0.750$ $\sigma_1 = 0.002$
Supplier II: $\mu_2 = 0.749$ $\sigma_2 = 0.0018$
Supplier III: $\mu_3 = 0.751$ $\sigma_3 = 0.0015$
Rivets are needed with a length that deviates at most 0.005 from the nominal value. Which supplier scores best in this respect?

4.52 Let $X \sim N(\mu, \sigma^2)$. Express the value a in terms of μ and σ, when
a $P(X > a) = 0.9$
b $P(X > a) = P(X \leq a) / 3$

4.53 Let $X \sim N(100, 5^2)$. There is a loss of €20 when the outcome of X is below 95, there is a profit of € 10 when $X > 110$, but the values in between give the highest profit, namely €50. Find the expected profit.

4.54 Rivets are useful only if the size X is between 0.25 and 0.38.
a Manufacturer A produces rivets with $N(0.30, 0.03^2)$-distributed sizes. What part is useful?
b If one could adjust the expected value of the size without affecting the standard deviation, for what μ will the fraction of useful rivets be as big as possible?

4.55 Let $Z \sim N(0, 1^2)$, thus standard normally distributed.
a Find $P(|Z| < 1.5)$.
b Find $P(Z^2 > 1)$.

4.56 Capacitors of type 'thingamagic' have capacity $X \sim N(5, 0.4^2)$. Four capacitors are needed with capacity between 4.3 and 5.9. Calculate the probability that four or five useful ones are included in a package of five.

4.57 The diameter of a cable is $X \sim N(0.8, 0.02^2)$.
a Calculate the probability of a diameter greater than 0.81.
b The cable is unusable if the diameter deviates more than 0.025 from the expected value. Find the probability of an unusable cable.
c Suppose the production process can be altered in such a way that the expected value remains constant at 0.8, while the standard deviation becomes smaller. Calculate the value of the standard deviation such that the fraction of unusable cables will be reduced to 0.10.

4.58 A one-pound bag of flour contains a $N(\mu, 0.02^2)$-distributed weight in pounds.
a For which μ will the probability of underweight (below 1 pound) be equal to 5%.
b In a batch 1000 bags with this μ are filled with a cost price of € 1. Bags with a minimal weight of 1 pound yield € 1.50, bags with underweight yield nothing. Find the expected profit of this batch.

4.59 Use the moment generating function to find the variance of a normal distribution with parameters μ and σ^2.

4.60 Determine $E(X^3)$ of a normal distribution as a function of μ and σ.

4.61 Consider draws (observations) from a $N(100, 20^2)$-distribution.
a Calculate the probability of an observation above 125.
b Calculate the probability that the sum of two observations turns out above 250.
c Calculate the probability that the average of 10 observations turns out above 125.

4.62 Consider two independent draws, both taken from a normal distribution with $\mu = 100$, but with σ equal to 15 and 25 respectively. Calculate the probability that the sum these two observations turns out above 250.

4.63 A random sample of 100 observations is taken from a UNIF(0, 1)-distribution.
a Calculate the probability that the sum of the 100 observations will be over 50.
b Approximate the probability that the sum of the 100 observations will be over 51.
c Approximate the probability that the *average* of the 100 observations turns out between 48.3 and 51.4.

4.64 Consider a distribution on the interval (0, 1) with the density function $2x$.
a Find the median m, the expected value μ and the standard deviation σ of this distribution.
b Find the probability that a random value from this distribution lies between μ and m.
c Approximate the probability that the sum of 100 draws from this distribution will be between 100μ and $100m$.

4.65 For each of the following probabilities, compare the exact value with the normal approximation (with cont.corr.):
a $P(X = 36 \mid X \sim \text{BIN}(100, 0.4))$
b $P(X = 18 \mid X \sim \text{BIN}(100, 0.2))$
c $P(X = 18 \mid X \sim \text{BIN}(50, 0.4))$
d $P(X = 9 \mid X \sim \text{BIN}(50, 0.2))$

4.66 Approximate the probability that when an unbiased die is thrown 600 times you get at least 110 times a six.

4.67 Determine the probabilities below, first by using a calculator or a computer, followed by normal approximation with continuity correction:
a $P(X \geq 7 \mid X \sim \text{BIN}(100, 0.05))$ (find this also with a table)
b $P(X \geq 324 \mid X \sim \text{BIN}(947, 0.34))$

4.68 Someone throws a fair coin 100 times. In this exercise always use the normal approximation for the binomial distribution.
a Show that the most likely outcome is 50 heads and 50 tails.
b Compare the probability of this outcome with the probability of 58 heads, and also with the probability of 66 heads.
c Find the set of all possible (elementary) outcomes with a probability that is 10 times as small or smaller than the probability of 50 heads and 50 tails.
d Find the total probability of this set.

4.69 A producer notes that 10% of his production is of B-quality. Calculate for a sample of 500 items from this production
a the probability that it contains 53 items of B-quality;
b the probability of at least 36 but at most 42 items of B-quality.

4.70 A system of 50 components functions 'reasonably' if at least 45 of these 50 are functioning properly.
a Given that the probability that a random component is defective equals $p = 0.15$, approximate the probability that the system functions reasonably. Assume that the occurrence of a failure of a component is independent of any failure of all the other components.
b What is the maximum value of p such that the probability of a reasonably functioning system is at least 0.95?

4.71 Approximate the following probabilities of the Poisson-distributed variables X, U and Y :
a $P(X > 60 \mid X \sim \text{POI}(64))$
b $P(U \leq 15 \mid U \sim \text{POI}(25))$
c $P(130 \leq Y \leq 161 \mid Y \sim \text{POI}(144))$

4.72 The number of seeds of a certain plant that falls on a test field in each month of October has a Poisson(35)-distribution. Approximate (with cont.corr.) the probability that next year October this number will be at least 25.

4.73 Since a GAM-distributed variable can be regarded as the sum of r exponentially distributed variables (provided that r is an integer number), the CLT is applicable and hence probabilities can be approximated with the normal distribution (if r is big enough). Let $r = 100$ and $\lambda = \theta^{-1} = 0.2$. Approximate the probability of an outcome above 505.

4.74 Use moment generating functions to prove that the sum of two independent chi-square distributed variables (with parameters m and n resp.), also has a chi-square distribution. With which parameter value?

4.75 Let $Z \sim \text{N}(0, 1^2)$, so standard normally distributed. Write the next probabilities in terms of a χ^2-distribution.
a $P(|Z| < 1.5)$
b $P(Z^2 > 1)$

4.76 Given that $Z \sim N(0, 1)$, calculate $P(Z^2 < 2)$ in two different ways.

4.77 Prove using an integral that $\text{E}(X) = \nu$, where $X \sim \chi^2(\nu)$.

4.78 Show that the mode of the chi-square distribution (with $\nu > 2$) is equal to $\nu - 2$.

4.79 Since a χ^2-distributed variable with ν degrees of freedom can be regarded as the sum of ν random variables (each chi-square distributed with 1 degree of freedom), the CLT is applicable. Hence, probabilities can be approximated with the normal distribution when ν is large enough.
- a Let $\nu = 100$. Approximate the probability of an outcome above 105.
- b Express the approximation of the probability of an outcome between the mode $(\nu - 2)$ and the expected value (ν) in terms of the CDF of the standard normal distribution (Φ).
- c Compare the answers of a and of b with the exact probability for a few cases selected by yourself (use a computer)

4.80 From the graph of the pdf of the chi-square distribution it seems like the density for $\nu = 3$ (resp. 1) intersects the density for $\nu = 5$ (resp. 3) exactly at its top. Generalize this thought and prove it.

4.81 Demonstrate how the probabilities of exercise 4.45 can be written as probabilities of chi-square distributed random variables.

4.82 The lifetime (in years) of a machine part has a Weibull-distribution with $\theta = 10$ and $\beta = 2$.
- a Calculate the probability that the part works at least three, but at most seven years.
- b Calculate the probability that a three year old part works at least four more years.

4.83 Let $U \sim$ UNIF(0,1). Use the CDF-method to find the density of:
- a $V = \sqrt{U}$
- b $W = \ln U$
- c $X = 1 - e^{-U}$
- d $Y = 1 / (U + 1)$
- e $Z = U(1 - U)$

4.84 When X is uniform on [2 , 5], find the density of $Y = 3X - 6$.

4.85 When X has density $f(x) = 2x$ on [0 , 1], find the density of $Y = X^2$.

4.86
- a When X is uniform on [−2 , 2] find the density of $Y = X^2$.
- b Find the expected value of Y using the density that was found in a.
- c It is known that the variance of the uniform distribution equals $b^2 / 12$, where b is the width of the support set (here: 4). How does the answer of part b follow from this?

4.87 When X is uniform on [−1 , 3], find the density of $Y = X^2$.

4.88 When X is uniform on [−1 , 1], find the density of $Y = X^3$.

4.89 Let a random variable X be uniformly distributed on the interval (−2 , 2).
Find the CDF and the pdf of $Y = 1 / X$.

4.90 When X on [0 , 1] has the density $4x(1 - x^2)$, then find
- a the expected value of X^2;
- b the density of $Y = X^2$ using the transformation method. Use the obtained density to find the expected value of Y, which is again the expected value of X^2.

4.91 Let X be uniformly distributed on [½ , 1]
- a Find the density of $Y = -\ln X$.
- b Is the following equality valid: $E(-\ln X) = -\ln E(X)$? What does Jensen's inequality say about this?

4.92 (B&E. 6.1) Let X be a continuous random variable with pdf $f_X(x) = 4x^3$ for $0 < x < 1$. Find the density function of the following random variables (use the CDF method as well as the transformation method):
- a $Y = X^4$
- b $W = e^X$
- c $Z = \ln X$
- d $U = (X - 0.5)^2$

4.93 (B&E, 6.4) Let $X \sim$ WEI(θ, β). Find the CDF and the pdf of:
- a $Y = (X/\theta)^\beta$
- b $W = \ln X$
- c $Z = (\ln X)^2$

4.94 When X has a continuous distribution on $[a\ ,\ b]$ with median m, and h is a strictly monotonous function, prove that the median of $Y = h(X)$ is equal to $h(m)$.

4.95 When X is uniform on $[-1\ ,\ 1]$, find a and b such that $Y = aX + b$ is uniform on $[0\ ,\ 100]$.

4.96 Answer the next questions using the CDF method as well as the transformation method.
- a When X has $f(x) = 2x$ on $[0\ ,\ 1]$, find the density of $Y = 4 - 4X$.
- b When X on $[0\ ,\ 1]$ has density $4x(1 - x^2)$, find the density of $Y = 10 - 5X$.
- c When X on $[0\ ,\ \infty)$ has an exponential distribution with expected value θ, find the density of $Y = 2X + 10$.

4.97 When X is continuous with expected value μ and standard deviation σ, prove that $Y = (X - \mu) / \sigma$ has expected value 0 and variance 1.

4.98 (B&E, Ex. 6.3.7) The discrete random variable X has probability distribution function $f_X(x) = \frac{4}{31}\left(\frac{1}{2}\right)^x$ for $x = -2, -1, 0, 1, 2$. Use the transformation method to determine the probability function for $Y = |X|$.

4.99 Let X have a normal $N(\mu\ ,\ \sigma^2)$-distribution.
- a Find the density of $Y = |X|$.
- b Find $E(|X|)$ in case $\mu = 0$.

4.100 Give the inverse of the cumulative distribution function of
- a the exponential distribution with expected value μ;
- b the uniform distribution on (a , b);
- c the Weibull distribution with parameters θ and β.

4.101 (B&E. 6.6) Let X be a continuous random variable with density $f_X(x) = 4x^3$ for $0 < x < 1$. Find the transformation $y = g(x)$ such that $Y = g(X) \sim \text{UNIF}(0,1)$.

4.102 When U is a UNIF(0 , 1)-variable, then $1 - U$ has the same distribution. Let $F(\ )$ be a function that satisfies the requirements of a CDF.
- a Do $X = F^{-1}(U)$ and $Y = F^{-1}(1-U)$ have the same distribution? Are X and Y identical?
- b For which F would it hold that $X = -Y$ for all possible outcomes of U?

4.103 When people are not able or not willing to make a decision, they sometimes toss a coin and let their choice depend on the outcome Heads or Tails. Describe how a random number generator (which generates numbers between 0 and 1) could be used, if someone has to choose between three options, where option 1 must have a probability of 20%, option 2 a probability of 45% and option 3 a probability of 35%.

4.104 Suppose it must be investigated whether a series of observations $x_1, x_2, x_3, \ldots, x_n$, in ascending order, could possibly be drawn from a distribution with CDF F. If a graph is made of

$$x_i \quad \text{against} \quad F^{-1}\left(\frac{i}{n+1}\right)$$

what picture would you get (approximately)?

4.105 Given is the following probability density function of the continuous random variable X.

$f_X(x) = x^2 \qquad \text{for } 0 < x < a$

- a Find the cumulative distribution function of X.
- b Derive the value of a using the CDF of part a.
- c We intend to simulate numbers that should come from the above distribution. The computer has a 'random number generator' that generates numbers that are uniformly distributed between 0 and 1. How can we transform these numbers such that they can be regarded as draws from the above distribution?

4.106 For this exercise, you should use Excel. In the first column, draw one hundred numbers with the standard Excel-function RANDOM(). Next, transform these with a suitable operation in column 2 to numbers that can be regarded as draws from an exponential distribution with expected value 3. Put these 100 numbers in ascending order. Which part of the sample is smaller than 2? Compare this with the probability $P(X < 2 \mid X \sim \text{EXP}(1/3))$. Is it approximately the same?

4.107 (B&E, 8.17) Used tabled values to find the following probabilities:

a $P(7.26 < Y < 22.31)$ where $Y \sim \chi^2(15)$

b The value of b such that $P(Y < b) = 0.9$ where $Y \sim \chi^2(23)$

c $P\left[\frac{Y}{1+Y} > \frac{11}{16}\right]$ where $Y \sim \chi^2(6)$

4.108 The hazard rate (see Example 4.11) of a chip is: $h(t) = \frac{f(t)}{R(t)} = \frac{t^2}{10^6}$

a Find the reliability $R(t)$, and sketch it.
(Hint: $R(t)$ can often be found using the following result for a given hazard rate: From the relation $R(t) = 1 - F(t)$ it follows of course that $R'(t) = -f(t)$. With this we can derive:
$h(t) = \frac{f(t)}{R(t)} = \frac{-R'(t)}{R(t)} = -\frac{d}{dt}\left(\ln R(t)\right)$)

b Sketch the density.

c Calculate the probability that the chip's time to failure is less than 100 hours.

d If n such chips are connected in parallel, and the device, in which the chips are placed, functions if at least one chip is working, then find the minimal number of chips that should be connected in parallel in order to get a probability of at least 0.99 that the device will reach a service time of 100 hours.

Appendix A

A.1 Summation and product signs

A sum of a number of terms can be written using the sum sign.

$$\sum_{i=2}^{5} i^2 = 2^2 + 3^2 + 4^2 + 5^2 \qquad \sum_{i=1}^{n} x_i = x_1 + x_2 + \ldots + x_n \qquad \sum_{i=1}^{15} 4 = 4 + 4 + \ldots + 4 + 4 = 15 \times 4 = 60$$

A product of factors can be written using the product sign.

$$\prod_{i=2}^{5} i^2 = 2^2 \times 3^2 \times 4^2 \times 5^2 = \left(\prod_{i=2}^{5} i \right)^2 = 14400 \qquad \prod_{i=1}^{n} x_i = x_1 \times x_2 \times \ldots \times x_{n-1} \times x_n = x_1 x_2 \ldots x_{n-1} x_n$$

Note:

$$\log(\prod_{i=1}^{n} x_i) = \log(x_1 x_2 \ldots x_n) = \log x_1 + \log x_2 + \ldots + \log x_n = \sum_{i=1}^{n} \log(x_i)$$

$$\prod_{i=1}^{n} e^{x_i} = e^{x_1} \times e^{x_2} \times \ldots \times e^{x_n} = e^{x_1 + x_2 + \ldots + x_n} = e^{\left(\sum_{i=1}^{n} x_i \right)} = \exp\left(\sum_{i=1}^{n} x_i \right)$$

A.2 Partial integration

$$\int f(x)\, dg(x) = f(x)\, g(x) - \int g(x)\, df(x) \qquad \text{where:} \quad dg(x) = g'(x)dx \quad \text{and} \quad df(x) = f'(x)dx$$

Can be remembered by using the product rule for differentiating:

$$f(x)\, g'(x) = \frac{d}{dx}\big(f(x)\, g(x) \big) \quad - \quad g(x)\, f'(x)$$

Another formulation for partial integration is:

$$\int u(x)\, v(x)\, dx = u(x)\, V(x) - \int V(x)\, u'(x)\, dx \qquad \text{where:} \quad \int v(x)\, dx = V(x)$$

A.3 Series

$$\sum_{i=1}^{n} i = \frac{n(n+1)}{2}$$

$$\sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}$$

$$\sum_{i=0}^{n} r^i = \frac{1 - r^{n+1}}{1 - r}$$

$$\sum_{i=0}^{\infty} r^i = \frac{1}{1 - r} \quad \text{when } |r| < 1$$

$$\sum_{i=1}^{\infty} i r^i = \frac{r}{(1 - r)^2} \quad \text{when } |r| < 1$$

$$\sum_{i=0}^{\infty} \frac{x^i}{i!} = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \ldots. = e^x$$

A.4 The Γ-function

For any real number $x > 0$, the gamma function is defined by:

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t}\, dt$$

By simple integration, we find:

$$\Gamma(1) = \int_0^\infty t^{1-1} e^{-t}\, dt = \int_0^\infty t^0\, d(-e^{-t}) = \left[-e^{-t}\right]_0^\infty = 1$$

By partial integration, we obtain:

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t}\, dt = \int_0^\infty t^{x-1}\, d(-e^{-t}) = \left[t^{x-1}(-e^{-t})\right]_0^\infty - \int_0^\infty (-e^{-t})\, d(t^{x-1}) =$$

$$= 0 + \int_0^\infty (x-1)\, t^{x-2} e^{-t}\, dt = (x-1) \int_0^\infty t^{(x-1)-1}\, e^{-t}\, dt = (x-1)\Gamma(x-1)$$

So for integer n: $\Gamma(n) = (n-1)\,\Gamma(n-1) = (n-1)\,(n-2)\,\Gamma(n-2) = (n-1)\,(n-2)\,\ldots\,\Gamma(1) = (n-1)!$

Also (without proof): $\Gamma(\tfrac{1}{2}) = \sqrt{\pi} \;\Rightarrow\; \Gamma(1\tfrac{1}{2}) = \tfrac{1}{2}\sqrt{\pi} \;\Rightarrow\; \Gamma(2\tfrac{1}{2}) = \tfrac{3}{4}\sqrt{\pi}$.

For $n!$ the following approximation is useful: $n! \approx \left(\frac{n}{e}\right)^n \sqrt{2\pi n}$ (*Stirling's Formula*)

A.5 Greek alphabet

A α alfa	H η èta	N ν nu	T τ tau
B β bèta	Θ θ thèta	Ξ ξ xi	Y υ ypsilon
Γ γ gamma	I ι iota	O ο omikron	Φ φ phi
Δ δ delta	K κ kappa	Π π pi	X χ chi
E ε epsilon	Λ λ lambda	P ρ rho	Ψ ψ psi
Z ζ zèta	M μ mu	Σ σ sigma	Ω ω omega

Appendix B

Table 1. Binomial Distribution

Tabled is: $P(X \le k) = \sum_{x=0}^{k} p(x)$.

***n* = 5**

	p														
k	**.01**	**.05**	**.10**	**.20**	**.25**	**.30**	**.40**	**.50**	**.60**	**.70**	**.75**	**.80**	**.90**	**.95**	**.99**
0	.951	.774	.590	.328	.237	.168	.078	.031	.010	.002	.001	.000	.000	.000	.000
1	.999	.977	.919	.737	.633	.528	.337	.188	.087	.031	.016	.007	.000	.000	.000
2	1.000	.999	.991	.942	.896	.837	.683	.500	.317	.163	.104	.058	.009	.001	.000
3	1.000	1.000	1.000	.993	.984	.969	.913	.813	.663	.472	.367	.263	.081	.023	.001
4	1.000	1.000	1.000	1.000	.999	.998	.990	.969	.922	.832	.763	.672	.410	.226	.049

***n* = 6**

	p														
k	**.01**	**.05**	**.10**	**.20**	**.25**	**.30**	**.40**	**.50**	**.60**	**.70**	**.75**	**.80**	**.90**	**.95**	**.99**
0	.941	.735	.531	.262	.178	.118	.047	.016	.004	.001	.000	.000	.000	.000	.000
1	.999	.967	.886	.655	.534	.420	.233	.109	.041	.011	.005	.002	.000	.000	.000
2	1.000	.998	.984	.901	.831	.744	.544	.344	.179	.070	.038	.017	.001	.000	.000
3	1.000	1.000	.999	.983	.962	.930	.821	.656	.456	.256	.169	.099	.016	.002	.000
4	1.000	1.000	1.000	.998	.995	.989	.959	.891	.767	.580	.466	.345	.114	.033	.001
5	1.000	1.000	1.000	1.000	1.000	.999	.996	.984	.953	.882	.822	.738	.469	.265	.059

***n* = 7**

	p														
k	**.01**	**.05**	**.10**	**.20**	**.25**	**.30**	**.40**	**.50**	**.60**	**.70**	**.75**	**.80**	**.90**	**.95**	**.99**
0	.932	.698	.478	.210	.133	.082	.028	.008	.002	.000	.000	.000	.000	.000	.000
1	.998	.956	.850	.577	.445	.329	.159	.063	.019	.004	.001	.000	.000	.000	.000
2	1.000	.996	.974	.852	.756	.647	.420	.227	.096	.029	.013	.005	.000	.000	.000
3	1.000	1.000	.997	.967	.929	.874	.710	.500	.290	.126	.071	.033	.003	.000	.000
4	1.000	1.000	1.000	.995	.987	.971	.904	.773	.580	.353	.244	.148	.026	.004	.000
5	1.000	1.000	1.000	1.000	.999	.996	.981	.938	.841	.671	.555	.423	.150	.044	.002
6	1.000	1.000	1.000	1.000	1.000	1.000	.998	.992	.972	.918	.867	.790	.522	.302	.068

***n* = 8**

	p														
k	**.01**	**.05**	**.10**	**.20**	**.25**	**.30**	**.40**	**.50**	**.60**	**.70**	**.75**	**.80**	**.90**	**.95**	**.99**
0	.923	.663	.430	.168	.100	.058	.017	.004	.001	.000	.000	.000	.000	.000	.000
1	.997	.943	.813	.503	.367	.255	.106	.035	.009	.001	.000	.000	.000	.000	.000
2	1.000	.994	.962	.797	.679	.552	.315	.145	.050	.011	.004	.001	.000	.000	.000
3	1.000	1.000	.995	.944	.886	.806	.594	.363	.174	.058	.027	.010	.000	.000	.000
4	1.000	1.000	1.000	.990	.973	.942	.826	.637	.406	.194	.114	.056	.005	.000	.000
5	1.000	1.000	1.000	.999	.996	.989	.950	.855	.685	.448	.321	.203	.038	.006	.000
6	1.000	1.000	1.000	1.000	1.000	.999	.991	.965	.894	.745	.633	.497	.187	.057	.003
7	1.000	1.000	1.000	1.000	1.000	1.000	.999	.996	.983	.942	.900	.832	.570	.337	.077

Binomial Distribution (continued)

Tabled is: $P(X \le k) = \sum_{x=0}^{k} p(x)$.

***n* = 9**

	p														
k	**.01**	**.05**	**.10**	**.20**	**.25**	**.30**	**.40**	**.50**	**.60**	**.70**	**.75**	**.80**	**.90**	**.95**	**.99**
0	.914	.630	.387	.134	.075	.040	.010	.002	.000	.000	.000	.000	.000	.000	.000
1	.997	.929	.775	.436	.300	.196	.071	.020	.004	.000	.000	.000	.000	.000	.000
2	1.000	.992	.947	.738	.601	.463	.232	.090	.025	.004	.001	.000	.000	.000	.000
3	1.000	.999	.992	.914	.834	.730	.483	.254	.099	.025	.010	.003	.000	.000	.000
4	1.000	1.000	.999	.980	.951	.901	.733	.500	.267	.099	.049	.020	.001	.000	.000
5	1.000	1.000	1.000	.997	.990	.975	.901	.746	.517	.270	.166	.086	.008	.001	.000
6	1.000	1.000	1.000	1.000	.999	.996	.975	.910	.768	.537	.399	.262	.053	.008	.000
7	1.000	1.000	1.000	1.000	1.000	1.000	.996	.980	.929	.804	.700	.564	.225	.071	.003
8	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.998	.990	.960	.925	.866	.613	.370	.086

***n* = 10**

	p														
k	**.01**	**.05**	**.10**	**.20**	**.25**	**.30**	**.40**	**.50**	**.60**	**.70**	**.75**	**.80**	**.90**	**.95**	**.99**
0	.904	.599	.349	.107	.056	.028	.006	.001	.000	.000	.000	.000	.000	.000	.000
1	.996	.914	.736	.376	.244	.149	.046	.011	.002	.000	.000	.000	.000	.000	.000
2	1.000	.988	.930	.678	.526	.383	.167	.055	.012	.002	.000	.000	.000	.000	.000
3	1.000	.999	.987	.879	.776	.650	.382	.172	.055	.011	.004	.001	.000	.000	.000
4	1.000	1.000	.998	.967	.922	.850	.633	.377	.166	.047	.020	.006	.000	.000	.000
5	1.000	1.000	1.000	.994	.980	.953	.834	.623	.367	.150	.078	.033	.002	.000	.000
6	1.000	1.000	1.000	.999	.996	.989	.945	.828	.618	.350	.224	.121	.013	.001	.000
7	1.000	1.000	1.000	1.000	1.000	.998	.988	.945	.833	.617	.474	.322	.070	.012	.000
8	1.000	1.000	1.000	1.000	1.000	1.000	.998	.989	.954	.851	.756	.624	.264	.086	.004
9	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.999	.994	.972	.944	.893	.651	.401	.096

***n* = 15**

	p														
k	**.01**	**.05**	**.10**	**.20**	**.25**	**.30**	**.40**	**.50**	**.60**	**.70**	**.75**	**.80**	**.90**	**.95**	**.99**
0	.860	.463	.206	.035	.013	.005	.000	.000	.000	.000	.000	.000	.000	.000	.000
1	.990	.829	.549	.167	.080	.035	.005	.000	.000	.000	.000	.000	.000	.000	.000
2	1.000	.964	.816	.398	.236	.127	.027	.004	.000	.000	.000	.000	.000	.000	.000
3	1.000	.995	.944	.648	.461	.297	.091	.018	.002	.000	.000	.000	.000	.000	.000
4	1.000	.999	.987	.836	.686	.515	.217	.059	.009	.001	.000	.000	.000	.000	.000
5	1.000	1.000	.998	.939	.852	.722	.403	.151	.034	.004	.001	.000	.000	.000	.000
6	1.000	1.000	1.000	.982	.943	.869	.610	.304	.095	.015	.004	.001	.000	.000	.000
7	1.000	1.000	1.000	.996	.983	.950	.787	.500	.213	.050	.017	.004	.000	.000	.000
8	1.000	1.000	1.000	.999	.996	.985	.905	.696	.390	.131	.057	.018	.000	.000	.000
9	1.000	1.000	1.000	1.000	.999	.996	.966	.849	.597	.278	.148	.061	.002	.000	.000
10	1.000	1.000	1.000	1.000	1.000	.999	.991	.941	.783	.485	.314	.164	.013	.001	.000
11	1.000	1.000	1.000	1.000	1.000	1.000	.998	.982	.909	.703	.539	.352	.056	.005	.000
12	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.996	.973	.873	.764	.602	.184	.036	.000
13	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.995	.965	.920	.833	.451	.171	.010
14	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.995	.987	.965	.794	.537	.140
15	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Binomial Distribution (continued)

n = 20

	p														
k	.01	.05	.10	.20	.25	.30	.40	.50	.60	.70	.75	.80	.90	.95	.99
0	.818	.358	.122	.012	.003	.001	.000	.000	.000	.000	.000	.000	.000	.000	.000
1	.983	.736	.392	.069	.024	.008	.001	.000	.000	.000	.000	.000	.000	.000	.000
2	.999	.925	.677	.206	.091	.035	.004	.000	.000	.000	.000	.000	.000	.000	.000
3	1.000	.984	.867	.411	.225	.107	.016	.001	.000	.000	.000	.000	.000	.000	.000
4	1.000	.997	.957	.630	.415	.238	.051	.006	.000	.000	.000	.000	.000	.000	.000
5	1.000	1.000	.989	.804	.617	.416	.126	.021	.002	.000	.000	.000	.000	.000	.000
6	1.000	1.000	.998	.913	.786	.608	.250	.058	.006	.000	.000	.000	.000	.000	.000
7	1.000	1.000	1.000	.968	.898	.772	.416	.132	.021	.001	.000	.000	.000	.000	.000
8	1.000	1.000	1.000	.990	.959	.887	.596	.252	.057	.005	.001	.000	.000	.000	.000
9	1.000	1.000	1.000	.997	.986	.952	.755	.412	.128	.017	.004	.001	.000	.000	.000
10	1.000	1.000	1.000	.999	.996	.983	.872	.588	.245	.048	.014	.003	.000	.000	.000
11	1.000	1.000	1.000	1.000	.999	.995	.943	.748	.404	.113	.041	.010	.000	.000	.000
12	1.000	1.000	1.000	1.000	1.000	.999	.979	.868	.584	.228	.102	.032	.000	.000	.000
13	1.000	1.000	1.000	1.000	1.000	1.000	.994	.942	.750	.392	.214	.087	.002	.000	.000
14	1.000	1.000	1.000	1.000	1.000	1.000	.998	.979	.874	.584	.383	.196	.011	.000	.000
15	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.994	.949	.762	.585	.370	.043	.003	.000
16	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.999	.984	.893	.775	.589	.133	.016	.000
17	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.996	.965	.909	.794	.323	.075	.001
18	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.999	.992	.976	.931	.608	.264	.017
19	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.999	.997	.988	.878	.642	.182

n = 25

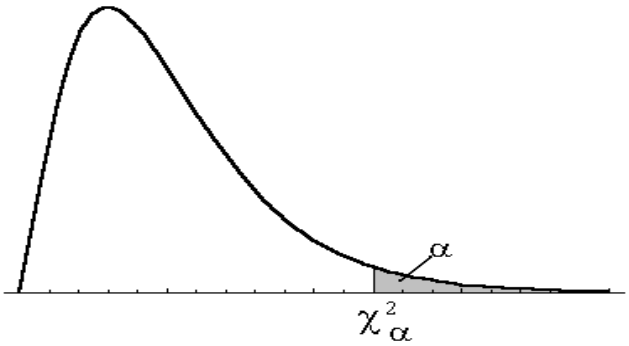
	p														
k	.01	.05	.10	.20	.25	.30	.40	.50	.60	.70	.75	.80	.90	.95	.99
0	.778	.277	.072	.004	.001	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
1	.974	.642	.271	.027	.007	.002	.000	.000	.000	.000	.000	.000	.000	.000	.000
2	.998	.873	.537	.098	.032	.009	.000	.000	.000	.000	.000	.000	.000	.000	.000
3	1.000	.966	.764	.234	.096	.033	.002	.000	.000	.000	.000	.000	.000	.000	.000
4	1.000	.993	.902	.421	.214	.090	.009	.000	.000	.000	.000	.000	.000	.000	.000
5	1.000	.999	.967	.617	.378	.193	.029	.002	.000	.000	.000	.000	.000	.000	.000
6	1.000	1.000	.991	.780	.561	.341	.074	.007	.000	.000	.000	.000	.000	.000	.000
7	1.000	1.000	.998	.891	.727	.512	.154	.022	.001	.000	.000	.000	.000	.000	.000
8	1.000	1.000	1.000	.953	.851	.677	.274	.054	.004	.000	.000	.000	.000	.000	.000
9	1.000	1.000	1.000	.983	.929	.811	.425	.115	.013	.000	.000	.000	.000	.000	.000
10	1.000	1.000	1.000	.994	.970	.902	.586	.212	.034	.002	.000	.000	.000	.000	.000
11	1.000	1.000	1.000	.998	.989	.956	.732	.345	.078	.006	.001	.000	.000	.000	.000
12	1.000	1.000	1.000	1.000	.997	.983	.846	.500	.154	.017	.003	.000	.000	.000	.000
13	1.000	1.000	1.000	1.000	.999	.994	.922	.655	.268	.044	.011	.002	.000	.000	.000
14	1.000	1.000	1.000	1.000	1.000	.998	.966	.788	.414	.098	.030	.006	.000	.000	.000
15	1.000	1.000	1.000	1.000	1.000	1.000	.987	.885	.575	.189	.071	.017	.000	.000	.000
16	1.000	1.000	1.000	1.000	1.000	1.000	.996	.946	.726	.323	.149	.047	.000	.000	.000
17	1.000	1.000	1.000	1.000	1.000	1.000	.999	.978	.846	.488	.273	.109	.002	.000	.000
18	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.993	.926	.659	.439	.220	.009	.000	.000
19	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.998	.971	.807	.622	.383	.033	.001	.000
20	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.991	.910	.786	.579	.098	.007	.000
21	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.998	.967	.904	.766	.236	.034	.000
22	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.991	.968	.902	.463	.127	.002
23	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.998	.993	.973	.729	.358	.026
24	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.999	.996	.928	.723	.222

Table 2. Poisson Distribution

Tabled is: $P(X \le k) = \sum_{x=0}^{k} p(x)$.

k	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0	6.0	7.0	8.0	9.0	10.0	11.0	12.0	13.0	14.0	15.0
0	.368	.223	.135	.082	.050	.030	.018	.011	.007	.002	.001	.000	.000	.000	.000	.000	.000	.000	.000
1	.736	.558	.406	.287	.199	.136	.092	.061	.040	.017	.007	.003	.001	.000	.000	.000	.000	.000	.000
2	.920	.809	.677	.544	.423	.321	.238	.174	.125	.062	.030	.014	.006	.003	.001	.001	.000	.000	.000
3	.981	.934	.857	.758	.647	.537	.433	.342	.265	.151	.082	.042	.021	.010	.005	.002	.001	.000	.000
4	.996	.981	.947	.891	.815	.725	.629	.532	.440	.285	.173	.100	.055	.029	.015	.008	.004	.002	.001
5	.999	.996	.983	.958	.916	.858	.785	.703	.616	.446	.301	.191	.116	.067	.038	.020	.011	.006	.003
6	1.000	.999	.995	.986	.966	.935	.889	.831	.762	.606	.450	.313	.207	.130	.079	.046	.026	.014	.008
7		1.000	.999	.996	.988	.973	.949	.913	.867	.744	.599	.453	.324	.220	.143	.090	.054	.032	.018
8			1.000	.999	.996	.990	.979	.960	.932	.847	.729	.593	.456	.333	.232	.155	.100	.062	.037
9				1.000	.999	.997	.992	.983	.968	.916	.830	.717	.587	.458	.341	.242	.166	.109	.070
10					1.000	.999	.997	.993	.986	.957	.901	.816	.706	.583	.460	.347	.252	.176	.118
11						1.000	.999	.998	.995	.980	.947	.888	.803	.697	.579	.462	.353	.260	.185
12							1.000	.999	.998	.991	.973	.936	.876	.792	.689	.576	.463	.358	.268
13								1.000	.999	.996	.987	.966	.926	.864	.781	.682	.573	.464	.363
14									1.000	.999	.994	.983	.959	.917	.854	.772	.675	.570	.466
15										.999	.998	.992	.978	.951	.907	.844	.764	.669	.568
16										1.000	.999	.996	.989	.973	.944	.899	.835	.756	.664
17											1.000	.998	.995	.986	.968	.937	.890	.827	.749
18												.999	.998	.993	.982	.963	.930	.883	.819
19												1.000	.999	.997	.991	.979	.957	.923	.875
20													1.000	.998	.995	.988	.975	.952	.917
21														.999	.998	.994	.986	.971	.947
22														1.000	.999	.997	.992	.983	.967
23															1.000	.999	.996	.991	.981
24																.999	.998	.995	.989
25																1.000	.999	.997	.994
26																	1.000	.999	.997
27																		.999	.998
28																		1.000	.999
29																			1.000

Table 3. Critical values for the chi-square distribution



Degrees of freedom	$\chi^2_{.995}$	$\chi^2_{.99}$	$\chi^2_{.975}$	$\chi^2_{.95}$	$\chi^2_{.90}$	$\chi^2_{.10}$	$\chi^2_{.05}$	$\chi^2_{.025}$	$\chi^2_{.01}$	$\chi^2_{.005}$
1	.0000393	.000157	.000982	.003932	.0158	2.706	3.841	5.024	6.635	7.879
2	.0100	.0201	.0506	.103	.211	4.605	5.991	7.378	9.210	10.597
3	.0717	.115	.216	.352	.584	6.251	7.815	9.348	11.345	12.838
4	.207	.297	.484	.711	1.064	7.779	9.488	11.143	13.277	14.860
5	.412	.554	.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	.676	.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.759	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.535	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.392	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.930	82.358	118.498	124.342	129.561	135.807	140.169

Table 4. Standard normal distribution: $\Phi(z) = \mathrm{P}(Z \le z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2}\, dt$

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998
3.5	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998
3.6	.9998	.9998	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999
3.7	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999
3.8	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999

Inverse standard normal distribution: $z_\alpha = \Phi^{-1}(1-\alpha)$

α	z_α
0.5	0.000
0.1	1.282
0.05	1.645
0.025	1.960
0.01	2.326
0.005	2.576
0.0025	2.807
0.001	3.090

Common distributions

Distribution	Pdf	Parameters	Support	Expectation	Variance	Mgf $E(e^{tX})$
Binomial(n , p)	$\binom{n}{x} p^x (1-p)^{n-x}$	n , p	{0, 1, 2, 3, …, n}	np	$np(1\text{-}p)$	$(pe^t + q)^n$
Negative-binomial(r , p)	$\binom{x-1}{r-1} p^r (1-p)^{x-r}$	p , r	{r, r+1, r+2, …}	$\frac{r}{p}$	$\frac{r(1-p)}{p^2}$	$\left(\frac{pe^t}{1-qe^t}\right)^r$
Geometric(p)	$p(1-p)^{x-1}$	p	{1, 2, 3, …}	$\frac{1}{p}$	$\frac{1-p}{p^2}$	$\frac{pe^t}{1-qe^t}$
Poisson(λ)	$e^{-\lambda}\frac{\lambda^x}{x!}$	λ	{0, 1, 2, 3, …}	λ	λ	$e^{\lambda(e^t-1)}$
Hypergeometric(n, M, N)	$\frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$	n M, N	$\max(0\,, M+n-N) \le$ $\le x \le \min(n\,, M)$	$\frac{nM}{N}$	$n\frac{M}{N}\frac{N-M}{N}\frac{N-n}{N-1}$	not useful
Uniform(a,b)	$\frac{1}{b-a}$	$a < b$	[a , b] of (a , b)	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	$\frac{e^{bt}-e^{at}}{(b-a)t}$
Normal(μ , σ^2)	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$	μ , σ^2 (of: σ)	($-\infty$, ∞)	μ	σ^2	$e^{\mu t+\sigma^2 t^2/2}$
Exponential(λ)	$\lambda e^{-\lambda x}$	$\lambda > 0$	[0 , ∞)	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	$\frac{\lambda}{\lambda - t}$
Gamma(θ, r)	$\frac{x^{r-1}e^{-x/\theta}}{\theta^r \Gamma(r)}$	θ , $r > 0$	[0 , ∞)	$r\theta$	$r\theta^2$	$\left(\frac{1}{1-\theta t}\right)^r$
Chi-square(ν)	$\frac{x^{\nu/2-1}e^{-x/2}}{2^{\nu/2}\Gamma(\nu/2)}$	$\nu = 1, 2, \ldots$	[0 , ∞)	ν	2ν	$(1-2t)^{-\nu/2}$