Probability Theory and Statistics 2

Luuk Seelen

Probability Theory and Statistics 2

Luuk Seelen

Acknowledgements: This text uses parts written originally by Hans Bührmann, who wrote a syllabus in the past for the Dutch language course ‘Kansrekening en Statistiek 2’. It also contains a number of problems which were taken from the book by Bain & Engelhardt, ‘Introduction to Probability and Mathematical Statistics’, 2nd edition. Furthermore, some minor parts were copied from Wikipedia. Finally, many remarks made by students in the past are reflected in this text as well.

Last update: March 29, 2023

5 Multivariate distributions

In the course Probability Theory and Statistics 1 we have been dealing with one-dimensional random variables. We will now consider situations where more than one random variable is defined within a single experiment, in which case a vector of random variables results. Thus, for example, in an experiment consisting of selecting at random a person from a population, we might define the random variables A (age), H (height), S (shoe size) and W (width), and put these into the 4-dimensional vector *(A, H, S, W)*. In general we will use the notation $\boldsymbol{X} = (X_1, ..., X_k)$ for a random vector which can take on the values $\boldsymbol{x} = (x_1, ..., x_k)$ in a k-dimensional space. The distribution of possible outcomes is called a **joint (probability) distribution** (or **multivariate distribution** (Dutch: simultane of gezamenlijke kansverdeling). Whenever $k = 2$, then we will usually use the term **bivariate distribution,** and use the notation X and Y, instead of X_1 and X_2. In this chapter, new concepts will first be explained using bivariate distributions, after which they will be generalised to the general multivariate situation. Just as before, we will make a distinction between discrete and continuous distributions.

5.1 Joint discrete distributions

(B&E, pages 139-144)

When the measuring scales for the two random variables X and Y are discrete, we can simply define the **joint probabilities** (Dutch: simultane kansen) as $P(X = x \cap Y = y) = P(X = x, Y = y)$.

Example 5.1

When rolling two unbiased dice, we can, for example, define X as the number of dots on the first die, and Y as the number of dots on the second. We can then create a **joint probability table** (cross-table) with the joint probabilities for each of the possible combinations of values that X and Y can attain. That table is very simple here, because $P(X = x, Y = y) = 1/36$ for $x = 1, 2, ..., 6$ and $y = 1, 2, ..., 6$.

		X: number of dots on first die					
		1	2	3	4	5	6
Y: number of dots on second die	1	1/36	1/36	1/36	1/36	1/36	1/36
	2	1/36	1/36	1/36	1/36	1/36	1/36
	3	1/36	1/36	1/36	1/36	1/36	1/36
	4	1/36	1/36	1/36	1/36	1/36	1/36
	5	1/36	1/36	1/36	1/36	1/36	1/36
	6	1/36	1/36	1/36	1/36	1/36	1/36

◄

Definition 5.1 *(B&E, Def. 4.2.1)*

The **joint probability distribution function** (or joint, simultaneous or bivariate pdf; Dutch: simultane kansfunctie) for the 2-dimensional *discrete* random variable (X, Y) is defined as:

$$f_{X,Y}(x, y) = P(X = x \cap Y = y) = P(X = x, Y = y)$$

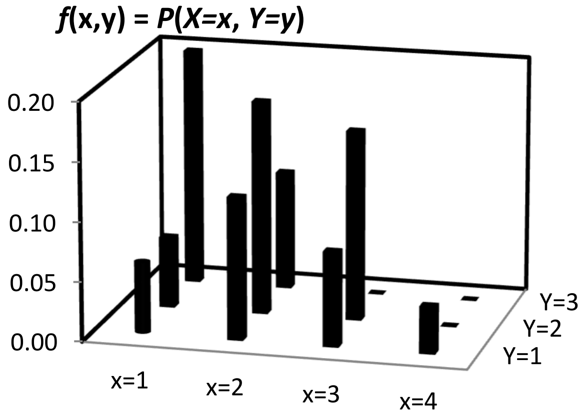
for all possible values (x, y) of (X, Y).

Example 5.2

A shop sells two different items A and B, and we define X and Y as the numbers of units sold of A and B respectively on any arbitrary day. The table below shows the joint probabilities; clearly those probabilities cannot be calculated in this case, but they may have been derived using historical sales data.

		X = Number of units A sold			
		0	1	2	3
Y = Number of units B sold	0	0.06	0.12	0.08	0.04
	1	0.06	0.18	0.16	0
	2	0.20	0.10	0	0

According to this table, $f_{X,Y}(1, 2) = 0.10$, or: the probability of selling on any arbitrary day exactly 1 unit of A and 2 units of B is equal to 0.10. This joint pdf may also be depicted graphically as follows:



◄

Analogous to univariate pdf's, each bivariate discrete pdf should possess the following (trivial) properties:

Theorem 5.1 *(B&E, Th.4.2.1)*

A function $f_{X,Y}(x,y)$ is the joint pdf for some two-dimensional discrete random variable (X,Y) if and only if:

$$f_{X,Y}(x,y) \geq 0 \quad \text{(for all } (x,y)\text{)} \quad \text{and} \quad \sum_{\text{all } x} \sum_{\text{all } y} f_{X,Y}(x,y) = 1$$

The univariate pdf's for X and Y can easily be derived from the joint pdf. For the case in Example 5.2, we can thus find the probability that $X = 1$ as follows (see also reader Prob. Theory and Statistics 1, section 2.5.3):

$$P(X=1) = P(X=1 \cap Y=0) + P(X=1 \cap Y=1) + P(X=1 \cap Y=2) = 0.40$$

Or, using another notation:

$$f_X(1) = f_{X,Y}(1,0) + f_{X,Y}(1,1) + f_{X,Y}(1,2) = \sum_{\text{all } y} f_{X,Y}(1,y) = 0.40 \, .$$

When a pdf has been derived from the joint pdf, it is usually called a *marginal* pdf:

Definition 5.2 (*B&E, Def. 4.2.2*)

If X and Y are discrete random variables with the joint pdf $f_{X,Y}(x,y)$, then the **marginal pdf's** of X and Y are:

$$f_X(x) = \sum_{\text{all } y} f_{X,Y}(x,y) \qquad \text{and} \qquad f_Y(y) = \sum_{\text{all } x} f_{X,Y}(x,y)$$

In a two-dimensional table with the joint probabilities, the marginal pdf's can be found by determining the row and column totals.

Example 5.3 (Example 5.2 continued)

		X = Number of units A sold				
		0	1	2	3	$f_Y(y)$
Y = Number of units B sold	0	0.06	0.12	0.08	0.04	0.30
	1	0.06	0.18	0.16	0	0.40
	2	0.20	0.10	0	0	0.30
$f_X(x)$		0.32	0.40	0.24	0.04	1.00

◄

Remark. These pdf's are often simply notated as $f(x)$, $f(y)$ and $f(x, y)$ (without subscripts). Note that this shorthand notation can be ambiguous as soon as the letters x or y are replaced by a number (or an expression); for example, does $f(3)$ refer to $f_X(3)$ or to $f_Y(3)$? In this reader this shorthand notation is therefore avoided.

The idea of summing joint probabilities can be generalised to the probability of an event A which is defined on the sample space of (X, Y) :

$$P(A) = P((X,Y) \in A) = \sum_{(x,y)\in A}\sum f_{X,Y}(x, y)$$

Example 5.4 (Example 5.1 continued)

Using the table of Example 5.1, we determine the probability that the sum of the dots on two unbiased dice is equal to 5:

$$P(X+Y=5) = \sum_{(x,y) \text{ with } x+y=5}\sum f_{X,Y}(x, y)$$

$$= f_{X,Y}(1,4) + f_{X,Y}(2,3) + f_{X,Y}(3,2) + f_{X,Y}(4,1) = 4*(1/36) = 1/9$$ ◄

All the information in the joint pdf can also be derived from the joint cumulative distribution function (CDF). The CDF will be defined in an analogous way as we did for a univariate random variable.

Definition 5.3 (*B&E, Def. 4.2.3*)

The **joint cumulative distribution function (CDF)** (Dutch: simultane (of gezamenlijke, of bivariate) verdelingsfunctie) of the 2-dimensional random variable (X,Y) is defined as:

$$F_{X,Y}(x, y) = P(X \le x,\ Y \le y)$$

Note that this definition is not restricted to discrete random variables, so we can use the same definition for continuous random variables, which will be discussed in the next section.

Example 5.5 (Example 5.2 continued)

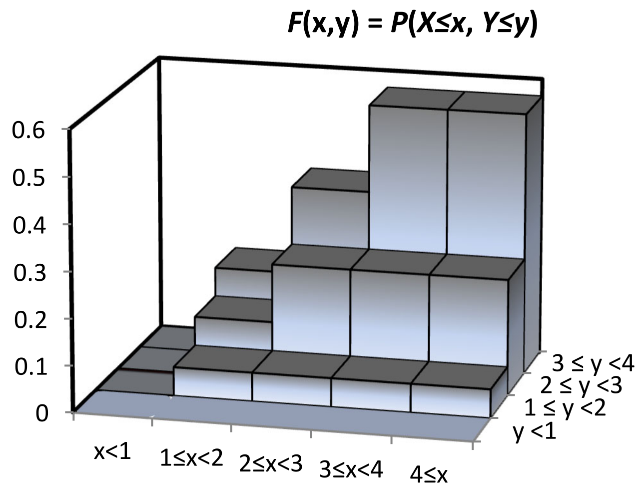
If we know the joint pdf, we can find the joint CDF by summing probabilities given by the pdf. Thus $F_{X,Y}(1,2) = P(X \le 1,\ Y \le 2) = 0.06 + 0.12 + 0.06 + 0.18 + 0.20 + 0.10 = 0.72$. But note that $F_{X,Y}(1.5, 2.6)$ is equal to 0.72 as well, just as for example $F_{X,Y}(1.999, 2)$ and $F_{X,Y}(1.0, 1000)$. In other words: $F_{X,Y}(x, y) = 0.72$ for all values of $x \in [1,\ 2)$ and $y \in [2,\ \infty)$! Note that such equality does not exist for the pdf, since for example $f_{X,Y}(1,2) = P(X=1,\ Y=2) = 0.10$, which is not equal to either $f_{X,Y}(1.5, 2.6)$ or $f_{X,Y}(1.999, 2)$ (since these last two probabilities are just 0). The joint CDF can be displayed in a table as follows:

	$x < 0$	$0 \le x < 1$	$1 \le x < 2$	$2 \le x < 3$	$3 \le x$
$y < 0$	0.00	0.00	0.00	0.00	0.00
$0 \le y < 1$	0.00	0.06	0.18	0.26	0.30
$1 \le y < 2$	0.00	0.12	0.42	0.66	0.70
$2 \le y$	0.00	0.32	0.72	0.96	1.00

Conversely, with some care, the joint pdf can also be deduced from the joint CDF, so for example:

$$f_{X,Y}(1, 2) = F_{X,Y}(1, 2) - F_{X,Y}(1, 1) - F_{X,Y}(0, 2) + F_{X,Y}(0, 1) \quad \text{(Check!)}$$

This CDF can be graphically displayed as follows:



◄

A joint CDF must meet certain conditions. In the bivariate situation, these are as follows:

Theorem 5.2 (*B&E, Th.4.2.2*)

A function $F_{X,Y}(x,y)$ is a bivariate CDF for a certain pair (*X, Y*) if and only if:

$$\lim_{x\to-\infty} F_{X,Y}(x,y) = F_{X,Y}(-\infty,y) = 0 \qquad \text{for all } y$$

$$\lim_{y\to-\infty} F_{X,Y}(x,y) = F_{X,Y}(x,-\infty) = 0 \qquad \text{for all } x$$

$$\lim_{x\to\infty, y\to\infty} F_{X,Y}(x,y) = F_{X,Y}(\infty,\infty) = 1$$

$$F_{X,Y}(b,d) - F_{X,Y}(b,c) - F_{X,Y}(a,d) + F_{X,Y}(a,c) \ge 0 \qquad \text{for all } a < b \text{ and } c < d$$

$$\lim_{h\downarrow 0} F_{X,Y}(x+h,y) = \lim_{h\downarrow 0} F_{X,Y}(x,y+h) = F_{X,Y}(x,y) \qquad \text{for all } x \text{ and } y$$

Proof

While the above might look complicated, these conditions follow all directly from the definition of a joint CDF. The fourth condition is apparent from the equation

$$P(a < X \le b,\ c < Y \le d) = F_{X,Y}(b,d) - F_{X,Y}(b,c) - F_{X,Y}(a,d) + F_{X,Y}(a,c)$$

and the fact that probabilities can never be negative. Note that this is a more stringent requirement than the condition that $F_{X,Y}(x,y)$ is non-decreasing in both of the two variables.

Example 5.6

Consider the function $F(x,y) = \begin{cases} 0 & \text{for } x+y<1 \\ 1 & \text{for } x+y \ge 1 \end{cases}$

This function cannot be a CDF, because the fourth condition is not satisfied. For example, choose $a = c = 0$ and $b = d = 1$, and we get:

$$P(0 < X \le 1,\ 0 < Y \le 1) = F(1,1) - F(1,0) - F(0,1) + F(0,0) = 1 - 1 - 1 + 0 = -1,$$

which of course is not possible. ◄

Multivariate discrete distributions

The concepts above for bivariate distributions can easily be generalised to *k*-dimensional vectors of discrete random variables.

Definition 5.4 (*B&E, Def. 4.2.1*)

The **joint pdf** of the *k*-dimensional discrete random variable $\boldsymbol{X} = (X_1,...,X_k)$ is defined as:

$$f_{X_1,...,X_k}(x_1,...,x_k) = P(X_1 = x_1,\ X_2 = x_2,...,X_k = x_k)$$

The notation $f_{X_1,...,X_k}(x_1,...,x_k)$ will often be shortened to $f_{\boldsymbol{X}}(x_1,...,x_k)$ or $f_{\boldsymbol{X}}(\boldsymbol{x})$.

Theorem 5.3 (*B&E, Th.4.2.1*)

A function $f_{X_1,...,X_k}(x_1,...,x_k)$ is the joint pdf for some *k*-dimensional discrete random variable $\boldsymbol{X} = (X_1,...,X_k)$ if and only if:

$$f_{X_1,...,X_k}(x_1,...,x_k) \geq 0 \quad \text{(for all } (x_1,...,x_k)) \qquad \text{and} \qquad \sum_{\text{all } x_1} \cdots \sum_{\text{all } x_k} f_{X_1,...,X_k}(x_1,...,x_k) = 1$$

Definition 5.5 (*B&E, Eq. 4.3.9*)

If the vector $(X_1,...,X_k)$ of discrete random variables has the joint pdf $f_{X_1,...,X_k}(x_1,...,x_k)$, then the **marginal pdf** of X_j (j = 1, …, k) is:

$$f_{X_j}(x_j) = \underset{\text{all } x_i \text{ with } i \neq j}{\sum \cdots \sum} f_{X_1,...,X_k}(x_1,...,x_k)$$

Definition 5.6 (*B&E, Def. 4.2.3*)

The **joint CDF** of the *k*-dimensional random variable $\boldsymbol{X} = (X_1,...,X_k)$ is defined as:

$$F_{X_1,...,X_k}(x_1,...,x_k) = P(X_1 \leq x_1,\ X_2 \leq x_2,...,X_k \leq x_k)$$

The multinomial distribution

In the course Prob. Theory and Statistics 1 we discussed the binomial distribution, which is the distribution of the number of successes *X* in a series of *n* independent trials, each with the same probability *p* of success. Each trial results in one of two possible outcomes (success or failure), where the probability of failure *q* is equal to 1 – *p*. Although there are two possible outcomes, the sample space for *X* is one-dimensional, as it is sufficient to count only the number of successes (or, if so desired, the number of failures).

The *multinomial distribution* is a generalisation of this binomial model: again we are dealing with *n* independent trials, but now each trial can result in *k* + 1 possible outcomes $E_1, \cdots, E_{k+1}$ (mutually exclusive), and where $p_i = P(E_i)$ denotes the probability that outcome E_i occurs. Since it is clear that $\sum_{i=1}^{k+1} p_i = 1$, it follows that $p_{k+1} = 1 - \sum_{i=1}^{k} p_i$. When we define X_i as the number of times E_i occurs, then it is obvious that $\sum_{i=1}^{k+1} X_i = n$, such that $X_{k+1} = n - \sum_{i=1}^{k} X_i$. Because X_{k+1} follows directly from $(X_1,...,X_k)$, the sample space is now k-dimensional (so not k +1-dimensional).

Example 5.7

When an unbiased die is rolled, there are six possible outcomes, so k+1 = 6 and $E_i = \{i \text{ dots}\}$ with $p_i = 1/6$. When we perform n trials of this experiment, we can count the number of times ($X_{1,\,...,}X_6$) that each of the categories E_1 to E_6 occur. We tally the number of 1s, the number of 2s, the number of 3s, etc. In this way, we find 6 results. But when we know the value of n, then we can deduce the number of 6s directly from $X_1,...,X_5$, since $X_6 = n - X_1 + X_2 + X_3 + X_4 + X_5$. This means that the sample space is not 6-dimensional, but only 5-dimensional, which also explains why k is equal to 5 here! ◄

In order to derive the formula for the multinomial pdf, we can proceed in a similar way as we did when we derived the binomial pdf in the previous course. We will do so here for the trinomial model, where each trial will result in one of three possible outcomes, A, B and C, with probabilities p_1, p_2, and p_3 respectively (such that $p_3 = 1 - p_1 - p_2$). Consider a sequence of outcomes for n trials. The result might be for example

A B B A C . . . (n letters in total)

When in this sequence we count x_1 times A, x_2 times B, and x_3 times C, then the probability of this sequence - *in this specific order* – to occur is equal to $p_1^{x_1} p_2^{x_2} p_3^{x_3}$. This same probability applies for all sequences with the same values for x_1, x_2 and x_3. In section 2.7.4 of the previous course, we discussed already how we can find the number of permutations in this case. Here, we will derive this number again, in a slightly different way. The x_1 letters A can be assigned to positions in the row in $\binom{n}{x_1}$ different ways. The remaining $n - x_1$ positions be filled with x_2 letters B and $x_3 = n - x_1 - x_2$ letters C. This can be done in $\binom{n-x_1}{x_2}$ ways. Therefore, the number of different sequences is:

$$\binom{n}{x_1} \times \binom{n-x_1}{x_2} = \frac{n!}{x_1!(n-x_1)!} \times \frac{(n-x_1)!}{x_2!(n-x_1-x_2)!} = \frac{n!}{x_1!x_2!x_3!} \qquad \text{(recall: } n - x_1 - x_2 = x_3\text{)}$$

This leads to the result:

$$f_{X_1,X_2}(x_1,x_2) = P(X_1 = x_1, X_2 = x_2) = P(X_1 = x_1, X_2 = x_2, X_3 = x_3) = \frac{n!}{x_1!x_2!x_3!} p_1^{x_1} p_2^{x_2} p_3^{x_3}$$

This idea can be easily generalised to the formula for the multinomial pdf:

Definition 5.7 (*B&E, Eq. 4.2.5*)

The vector $(X_1,...,X_k)$ of discrete random variables has a **multinomial** distribution, notation $\boldsymbol{X} \sim \text{MULT}(n, p_1,..., p_k)$, when the joint pdf is given by:

$$f_X(x_1,...,x_k) = \frac{n!}{x_1!x_2!...x_k!x_{k+1}!} p_1^{x_1} p_2^{x_2} ... p_{k+1}^{x_{k+1}} \qquad \text{for } x_i = 0,1,...,n$$

where $p_{k+1} = 1 - \sum_{i=1}^{k} p_i$, $p_i \geq 0$ $(i = 1,...,k+1)$ and $x_{k+1} = n - \sum_{i=1}^{k} x_i$.

Example 5.8

Throw 20 times with a fair die. The probability of obtaining exactly 2 'ones', 2 'twos', 2 'threes', 5 'fours' and 5 'fives' (*and therefore* 20 – 2 – 2 – 2 – 5 – 5 = 4 'sixes'!) is equal to:

$$f_X(2,2,2,5,5) = \frac{20!}{2!2!2!5!5!4!}\left(\frac{1}{6}\right)^{20} = 0.000241 \qquad ◄$$

What is the marginal pdf of X_j (with *j* between 1 and *k*) when $(X_1,...,X_k)$ has a multinomial distribution? Note that we are now only interested in the number of times the outcome E_j occurs, which could be defined as the number of successes, such that all other outcomes are seen as failures. It is clear then that the binomial distribution applies, so the marginal pdf should be:

$$f_{X_j}(x_j) = \binom{n}{x_j} p_j^{x_j}(1-p_j)^{n-x_j}$$

It might be instructive to formally prove this as well for *k*=2, using Definition 5.2 and/or Definition 5.5. If $(X_1, X_2) \sim \text{MULT}(n, p_1, p_2)$, then the marginal pdf of X_1 is:

$$f_{X_1}(x_1) = \sum_{x_2} f_{X_1,X_2}(x_1,x_2) = \sum_{x_2=0}^{n-x_1} \frac{n!}{x_1!x_2!x_3!} p_1^{x_1} p_2^{x_2} p_3^{x_3} =$$

$$= \frac{n!}{x_1!(n-x_1)!} p_1^{x_1} \sum_{x_2=0}^{n-x_1} \frac{(n-x_1)!}{x_2!(n-x_1-x_2)!} p_2^{x_2} p_3^{n-x_1-x_2} \qquad (\text{since } x_3 = n - x_1 - x_2)$$

$$= \binom{n}{x_1} p_1^{x_1} \sum_{x_2=0}^{n-x_1} \binom{n-x_1}{x_2} p_2^{x_2} p_3^{n-x_1-x_2}$$

$$= \binom{n}{x_1} p_1^{x_1} (p_2+p_3)^{n-x_1} \qquad (\text{applying the Binomial Theorem})$$

$$= \binom{n}{x_1} p_1^{x_1} (1-p_1)^{n-x_1} \qquad (x_1 = 0,1,...,n) \text{ so, indeed: } X_1 : \ Bin(n, p_1) \qquad ◄$$

The multivariate hypergeometric distribution

The hypergeometric distribution has been discussed in the previous course; it describes the distribution of the number of successes when sampling without replacement from a population with a finite number of elements, of which a fixed number are successes (and the rest failures). Now we assume that the population contains elements classified into more than two categories.

Example 5.9

A box contains 1000 plant seeds, of which 400 will result in red flowers, 400 in white flowers and 200 in pink flowers. Consider a random sample of size 10 (without replacement). We define X_1 as the number of seeds resulting in red flowers and X_2 as the number resulting in white flowers. Note that the number resulting in pink flowers will then be equal to $10 - X_1 - X_2$. Using the counting techniques discussed in chapter 2 (see previous course), the joint pdf of (X_1, X_2) can be found to be:

$$f_{X_1,X_2}(x_1,x_2) = \frac{\binom{400}{x_1}\binom{400}{x_2}\binom{200}{10-x_1-x_2}}{\binom{1000}{10}}$$

for $x_1 = 0,1,...,$ $x_2 = 0,1,...,$ and $x_1 + x_2 = 0,1,...,10$.

So the probability of selecting at random exactly 2 red, 5 white and 3 pink is

$$f_{X_1,X_2}(2,5) = \frac{\binom{400}{2}\binom{400}{5}\binom{200}{3}}{\binom{1000}{10}} = 0.0331$$

Note that we calculated similar probabilities already as part of the problems for chapter 2 of Prob. Theory and Statistics 1. But there, we did not formulate these in the form of a joint pdf. ◄

This example can be generalised as follows. Assume a set (population) with *N* items, which can be classified into *k* + 1 categories with M_i items belonging to category *i*, with $M_1 + M_2 + \ldots + M_{k+1} = N$. We select at random *n* items without replacement from this set, and we define X_i as the number of items of category *i* selected (for *i* = 1, … , *k* + 1):

Definition 5.8 (*B&E, Eq. 4.2.3*)

The vector $(X_1,\ldots,X_k)$ of discrete random variables has a **multivariate (or extended) hyper-geometric distribution**, notation $\boldsymbol{X} \sim \text{HYP}(n, M_1,\ldots,M_k, N)$, when the joint pdf is given by

$$f_{X_1,\ldots,X_k}(x_1,\ldots,x_k) = \frac{\binom{M_1}{x_1}\binom{M_2}{x_2}\cdots\binom{M_k}{x_k}\binom{M_{k+1}}{x_{k+1}}}{\binom{N}{n}} \qquad \text{for } x_i = 0,1,\ldots,M_i$$

where $M_{k+1} = N - \sum_{i=1}^{k} M_i$ and $x_{k+1} = n - \sum_{i=1}^{k} x_i$.

5.2 Joint continuous distributions

(B&E, pages 144-149)

The definition for the joint CDF (see Definition 5.2 and/or Definition 5.5) as well as Theorem 5.2 are equally valid for *continuous* random variables. Analogous to Definition 4.1 (previous course), we can define the bivariate pdf:

Definition 5.9 (*≈B&E, Def. 4.3.1*)

A 2-dimensional random variable (X,Y) is called continuous if a function $f_{X,Y}(\cdot,\cdot)$ exists, such that the CDF can be written as:

$$F_{X,Y}(x,y) = \int_{-\infty}^{y}\int_{-\infty}^{x} f_{X,Y}(u,v)\,du\,dv \qquad \left(= \int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(u,v)\,dv\,du\right)$$

We call $f_{X,Y}(x,y)$ the **joint probability density function (pdf)** (or bivariate, or simultaneous pdf, Dutch: simultane/gezamenlijke/bivariate kansdichtheid/dichtheidsfunctie).

Remark. See appendix A.6 for a short introduction to these double integrals. Multiple integrals can be evaluated starting from the inner integral. So

$$\int_a^b\int_c^d f(x,y)\,dy\,dx = \int_a^b\left[\int_c^d f(x,y)\,dy\right]dx .$$

The pdf can be derived from the joint CDF by differentiating (just like the univariate case), but now by first taking the partial derivative with respect to one variable and then again with respect to the other:

$$f_{X,Y}(x,y) = \frac{\partial^2 F_{X,Y}(x,y)}{\partial x\,\partial y} = \frac{\partial^2 F_{X,Y}(x,y)}{\partial y\,\partial x}$$

where we assume that the result does not change because of the different order in taking the partial derivatives (cases where that is the case are irrelevant for this discussion of multivariate probability theory). Recall that we are able to find probabilities in the univariate case by determining the area under the curve for the pdf; now we find probabilities by determining the relevant *volume* under the two-dimensional pdf. In general:

$$P(A) = P((X,Y) \in A) = \iint_{(x,y)\in A} f_{X,Y}(x,y)\,dx\,dy$$

For example:

$$P(a \le X \le b, c \le Y \le d) = \int_a^b\int_c^d f_{X,Y}(x,y)\,dy\,dx = \int_c^d\int_a^b f_{X,Y}(x,y)\,dx\,dy$$

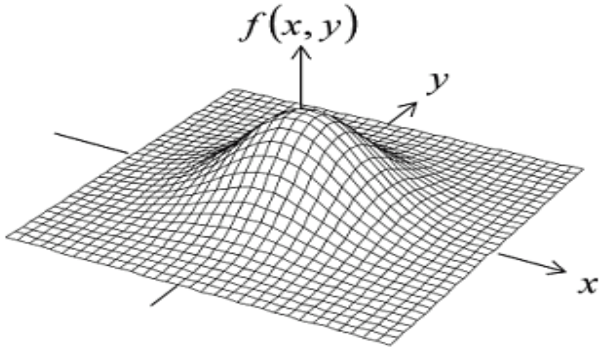
As a direct result, we get the following theorem:

Theorem 5.4 (*B&E, Th.4.3.1*)

A function $f_{X,Y}(x,y)$ is a bivariate pdf for a certain pair of continuous random variables (X,Y) if and only if:

$$f_{X,Y}(x,y) \ge 0 \quad \text{(for all } (x,y)\text{)} \quad \text{and} \quad \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx\,dy = 1$$

The following figure might represent a bivariate pdf, when the volume between the curve and the *xy*-plane is equal to 1:



The set of values for x and y that may occur as outcomes will again be called the support (set) of the distribution, formally written as (for the bivariate situation): $\{(x,y) \mid f_{X,Y}(x,y) > 0\}$. It should be emphasised that in general it is not sufficient to define a probability density by its formula; the support set should also be specified explicitly. The pdf will therefore be 0 for values of (x,y) not belonging to the support. Whenever in this reader an explicit function is given for a pdf with its associated support, we may implicitly assume that the pdf is equal to 0 elsewhere. When evaluating integrals and/or sums, it is essential to be well aware of the support set!

<u>*Example 5.10*</u>

Say the joint pdf for X and Y is given by:

$$f_{X,Y}(x,y) = \begin{cases} x^2 + \frac{8}{3}xy & \text{for } 0 < x \le 1 \text{ and } 0 < y \le 1 \\ 0 & \text{elsewhere} \end{cases}$$

The support set is here the set $\{(x,y) \mid 0 < x \le 1 \text{ en } 0 < y \le 1\}$. In this reader we will simply write:

$$f_{X,Y}(x,y) = x^2 + \frac{8}{3}xy \qquad \text{for } 0 < x \le 1 \text{ and } 0 < y \le 1$$

(Note it doesn't make any difference to write the support set as $0 \le x < 1$ and $0 < y < 1$ or $0 \le x \le 1$ and $0 \le y \le 1$, since the value for the pdf at singular points is irrelevant when we are dealing with continuous random variables).

That this is indeed a pdf is shown by the fact that the density is never negative, and that

$$\int_{-\infty}^{\infty}\int_{\infty}^{\infty} f_{X,Y}(x,y)\,dx\,dy = \int_0^1\int_0^1 (x^2 + \frac{8}{3}xy)\,dx\,dy = \int_0^1 \left(\frac{1}{3} + \frac{8}{6}y\right)dy = \frac{1}{3} + \frac{8}{12} = 1\,.$$

To find the joint CDF, we start with:

$$F_{X,Y}(x,y) = P(X \le x,\ Y \le y) = \int_{-\infty}^{y}\int_{-\infty}^{x} f_{X,Y}(u,v)\,du\,dv$$

Now, we must take into account various cases. For $0 \le x \le 1$ en $0 \le y \le 1$, we get:

$$F_{X,Y}(x,y) = \int_{-\infty}^{y}\int_{-\infty}^{x} f_{X,Y}(u,v)\,du\,dv = \int_0^{y}\int_0^{x}\left(u^2 + \frac{8}{3}uv\right)du\,dv$$

$$= \int_0^y \left(\frac{1}{3}u^3 + \frac{4}{3}u^2 v\right)\Bigg|_{u=0}^{u=x} dv = \int_0^y \left(\frac{1}{3}x^3 + \frac{4}{3}x^2 v\right)dv =$$

$$= \left(\frac{1}{3}x^3 v + \frac{2}{3}x^2 v^2\right)\Bigg|_{v=0}^{v=y} = \frac{1}{3}x^3 y + \frac{2}{3}x^2 y^2$$

But for $0 \le x \le 1$ and $y > 1$, it follows (note the upper limit of the first integral)

$$F_{X,Y}(x,y) = \int_0^1\int_0^x\left(u^2 + \frac{8}{3}uv\right)du\,dv = \frac{1}{3}x^3 + \frac{2}{3}x^2\,.$$

Of course, this result could have been found directly from the CDF for $0 \le x \le 1$ and $0 \le y \le 1$, because it is easy to see in this example that for all $0 \le x \le 1$ and $y > 1$, we obtain:

$$F_{X,Y}(x,y) = P(X \le x, Y \le y) = P(X \le x, Y \le 1) = F_{X,Y}(x,1)$$

So, by substituting in $\frac{1}{3}x^3 y + \frac{2}{3}x^2 y^2$ the variable y by the value 1, we obtain immediately the correct result.

Similarly, for $x > 1$ and $0 \le y \le 1$: $F_{X,Y}(x,y) = F_{X,Y}(1,y) = \frac{1}{3}y + \frac{2}{3}y^2$

For $x > 1$ and $y > 1$: $F_{X,Y}(x,y) = F_{X,Y}(1,1) = 1$ (this has to be 1, so this provides a good check!)

Finally for $x < 0$ or $y < 0$: $F_{X,Y}(x,y) = 0$ ◄

Example 5.11

For the 2-dimensional random variable from the previous example, we can compute the following probability:

$$P(X > 0.5,\ Y < 0.75) = \int_0^{0.75}\int_{0.5}^{1} (x^2 + \frac{8}{3}xy)\,dx\,dy$$

$$= \int_0^{0.75}\left(\frac{1}{3}x^3 + \frac{4}{3}x^2 y\right)\Bigg|_{x=0.5}^{x=1} dy = \int_0^{0.75}\left(\left(\frac{1}{3} - \frac{1}{24}\right) + \left(\frac{4}{3} - \frac{1}{3}\right)y\right)dy = \int_0^{0.75}\left(\frac{7}{24} + y\right)dy$$

$$= \left(\frac{7}{24}y + \frac{1}{2}y^2\right)\Bigg|_0^{0.75} = \frac{1}{2}$$

Of course, the result should not be dependent on the chosen integration order, as can be seen below:

$$P(X>0.5,\ Y<0.75)=\int_{0.5}^{1}\int_{0}^{0.75}(x^2+\frac{8}{3}xy)\,dy\,dx=\int_{0.5}^{1}\frac{3}{4}\left(x^2+x\right)dx=\frac{1}{2}$$

We can use the same idea for more 'complicated' probabilities as well, for example the probability that X is greater than Y:

$$\mathrm{P}(X>Y)=\int_0^1\int_y^1(x^2+\frac{8}{3}xy)\,dx\,dy$$

$$=\int_0^1\left(\frac{1}{3}x^3+\frac{4}{3}x^2y\right)\Bigg|_{x=y}^{x=1}dy=\int_0^1\left(\frac{1}{3}+\frac{4}{3}y-\frac{1}{3}y^3-\frac{4}{3}y^3\right)dy=$$

$$=\left(\frac{1}{3}y+\frac{2}{3}y^2-\frac{5}{12}y^4\right)\Bigg|_0^1=\frac{7}{12}$$

Always watch carefully the correct integration limits! When formulating the correct multiple integral, we work from the outside inwards. So in the case above: y can range from 0 to 1, but *given a specific value for y*, the variable x can range from y (because we are interested here in the probability that $X>Y$) to 1 (because a value for x greater than 1 would fall outside the support). When we start evaluating the resulting formula, we work from the inside out, as can be seen above.
It is useful to be aware that a probability such as $P(X>Y)$ should always result in a value between 0 and 1, and should never be a function of either x or y! Note that, on the other hand, $P(X>y)$ should in general result in a function of the value y!
By switching the order of integration, we get: $P(X>Y)=\int_0^1\int_0^x(x^2+\frac{8}{3}xy)\,dy\,dx$, of course giving the same result. ◄

Often, we are interested in the distribution of a single random variable, which will again be called a marginal distribution.

Definition 5.10 (*B&E, Def. 4.3.2*)

When the pair (X,Y) of continuous random variables has the joint pdf $f_{X,Y}(x,y)$, then the **marginal pdf's** of X and Y are:

$$f_X(x)=\int_{-\infty}^{\infty}f_{X,Y}(x,y)dy \qquad \text{and} \qquad f_Y(y)=\int_{-\infty}^{\infty}f_{X,Y}(x,y)dx$$

Note the similarity with the definition of a marginal pdf for discrete random variables. Also note that:

$$F_X(x)=P(X\le x)=P(X\le x,Y\le\infty)=\int_{-\infty}^{x}\int_{-\infty}^{\infty}f_{X,Y}(u,y)\,dy\,du$$

$$\Rightarrow f_X(x)=\int_{-\infty}^{\infty}f_{X,Y}(x,y)\,dy$$

(the last step because $f_X(x)=\dfrac{dF_X(x)}{dx}$ and the fact that for any continuous function $g(\cdot)$ it can be shown that $\dfrac{d\left[\int_{-\infty}^{x}g(u)du\right]}{dx}=g(x)$.

Example 5.12

Assume a probability distribution given by $f_{X,Y}(x,y) = (1-x)(2-y)$ for $(x, y) \in (0, 1) \times (0, 2)$.

This is a proper pdf, since the function value is never negative and we obtain the value 1 when we integrate this pdf over its support set:

$$\int_0^2 \int_0^1 (1-x)(2-y)dxdy = \int_0^2 \left[-\tfrac{1}{2}(1-x)^2(2-y)\right]_0^1 dy = \int_0^2 \left[\tfrac{1}{2}(2-y)\right]dy = \left[-\tfrac{1}{4}(2-y)^2\right]_0^2 = 1$$

The marginal pdf of X is found by integration over y:

$$f_X(x) = \int_0^2 (1-x)(2-y)dy = \left[-\tfrac{1}{2}(1-x)(2-y)^2\right]_0^2 = 2-2x \qquad \text{for } 0 \le x \le 1$$

And the marginal pdf of Y is:

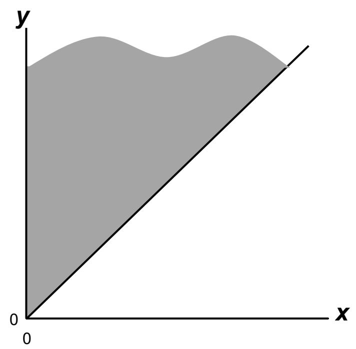
$$f_Y(y) = \int_0^1 (1-x)(2-y)dx = \left[-\tfrac{1}{2}(1-x)^2(2-y)\right]_0^1 = 1-\tfrac{1}{2}y \qquad \text{for } 0 \le y \le 2$$

If desired, one might check whether the resulting pdf's are indeed proper pdf's again (indeed, both are non-negative and the areas under the curves are equal to 1). ◄

Students often make mistakes when determining marginal pdf's. It might help to realise that the marginal pdf of X can only be a function of x, and the variable y should not occur in the function anymore (and vice versa). In order to check the result, one might check whether the resulting pdf's are indeed proper pdf's again (non-negative and an area under the curve equal to 1). Drawing a figure of the support set might also be helpful.

Example 5.13

Let $f_{X,Y}(x,y) = 2e^{-x-y}$ for $0 \le x \le y$. The *support set* can be displayed in a figure as follows:



The marginal pdf of X can be found by:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy$$

Since we want to determine the function value of the pdf for any arbitrary $x \ge 0$, we can treat x when we are evaluating the above integral as a given value (a constant). As soon as we replace $f_{X,Y}(x,y)$ by the formula $2e^{-x-y}$ in this integral, we must make sure that our integration limits for y take the support into account (because $f_{X,Y}(x,y)=0$ for values of (x, y) outside the support). So, given an arbitrary value for $x \ge 0$, we can see that y can range from x to infinity, Thus:

$$f_X(x) = \int_x^{\infty} 2e^{-x-y}dy = -2e^{-x-y}\Big|_{y=x}^{y=\infty} = 2e^{-2x} \qquad \text{for } x \ge 0$$

Similarly, the marginal pdf of Y will be (note that $0 \le x \le y$):

$$f_Y(y) = \int_0^y 2e^{-x-y}dx = -2e^{-x-y}\Big|_{x=0}^{x=y} = 2e^{-y}(1-e^{-y}) \qquad \text{for } y \ge 0 \text{ .}$$

Remark. We know that always $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx\,dy = 1$. When we would write this double integral for this example, we can find the limits of integration by working our way from the outside in: *y* can range from 0 to infinity, but given a value of *y, x* can then only range from 0 to *y*, with the result: $\int_0^{\infty}\int_0^{y} 2e^{-x-y}dx\,dy$. Here, the inner integral is exactly equal to the marginal pdf of *Y*, so

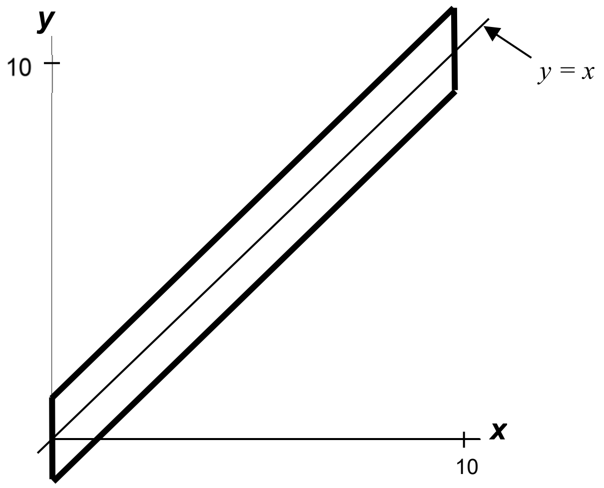
$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx\,dy = \int_0^{\infty}\int_0^{y} 2e^{-x-y}dx\,dy = \int_0^{\infty} f_Y(y)\,dy = 1\,.$$

Finally, with different order of integration:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy\,dx = \int_0^{\infty}\int_x^{\infty} 2e^{-x-y}dy\,dx = \int_0^{\infty} f_X(x)\,dx = 1$$ ◄

Example 5.14

Given is the joint pdf $f_{X,Y}(x,y) = 1/20$ for $0 < x < 10,\ x-1 < y < x+1$ (The support is concentrated around the line *y = x,* see figure below.



The determination of the marginal pdf of *X* is quite straightforward (it can help if you draw a vertical line for an arbitrary value of *x* between 0 and 10, and check which values for *y* fall within the support):

$$f_X(x) = \int_{x-1}^{x+1}\frac{1}{20}dy = \frac{1}{10}$$

But when determining the marginal pdf for *Y,* it becomes more complicated. For values of *y* between -1 and 1, *x* can only assume values between 0 and *y* + 1 (check this in the figure by drawing a horizontal line for any arbitrary value of $y \in (-1,\ 1)$). For values of *y* between 1 and 9, *x* can assume values between *y* – 1 and *y* + 1, and for $y \in (9,\ 11)$, *x* can assume values between *y* – 1 and 10. Therefore, we get:

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx = \begin{cases} \int_0^{y+1}\frac{1}{20}dx = \frac{y+1}{20} & \text{for } -1 < y < 1 \\ \int_{y-1}^{y+1}\frac{1}{20}dx = \frac{1}{10} & \text{for } 1 \le y \le 9 \\ \int_{y-1}^{10}\frac{1}{20}dx = \frac{11-y}{20} & \text{for } 9 < y < 11 \end{cases}$$

◄

Multivariate continuous distributions

We generalise the results of the previous paragraph to k-dimensional continuous random variables:

Definition 5.11 (*B&E, Def. 4.3.1*)

A k-dimensional random variable $\boldsymbol{X} = (X_1,...,X_k)$ is called continuous if a function $f_{X_1,...,X_k}(\cdot,...,\cdot)$ exists, such that the CDF can be written as:

$$F_{X_1,...,X_k}(x_1,...,x_k) = \int_{-\infty}^{x_k} \cdots \int_{-\infty}^{x_1} f_{X_1,...,X_k}(t_1,...,t_k)\, dt_1 \cdots dt_k$$

We call $f_{X_1,...,X_k}(x_1,...,x_k)$ a **joint or multivariate pdf**.

Then it follows: $$f_{X_1,...,X_k}(x_1,...,x_k) = \frac{\partial^k}{\partial x_1 \cdots \partial x_k} F_{X_1,...,X_k}(x_1,...,x_k)$$

And for any event A : $P((X_1,...,X_k) \in A) = \underset{(x_1,...,x_k)\in A}{\int \cdots \int} f_{X_1,...,X_k}(x_1,...,x_k)\, dx_1 \dots dx_k$

Theorem 5.5 (*B&E, Th.4.3.1*)

A function $f_{X_1,...,X_k}(x_1,...,x_k)$ is a joint pdf for some continuous random variables $(X_1,...,X_k)$ if and only if:

$$f_{X_1,...,X_k}(x_1,...,x_k) \geq 0 \quad \text{(for all } (x_1,...,x_k)\text{)} \quad \text{and}$$

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{X_1,...,X_k}(x_1,...,x_k)\, dx_1 \cdots dx_k = 1$$

Definition 5.12 (*B&E, Eq. 4.3.10*)

If the vector $(X_1,...,X_k)$ of continuous random variables has the joint pdf $f_{X_1,...,X_k}(x_1,...,x_k)$, then the **marginal pdf** of X_j (j = 1, …, k) is given by

$$f_{X_j}(x_j) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{X_1,...,X_k}(x_1,..,x_{j-1},x_j,x_{j+1},...,x_k)\, dx_1 \dots dx_{j-1} dx_{j+1} \dots dx_k$$

<u>*Example 5.15*</u>

Assume (X_1, X_2, X_3) to be continuous with pdf $f_{X_1,X_2,X_3}(x_1,x_2,x_3) = c$ for $0 < x_1 < x_2 < x_3 < 1$, where c is a certain constant. First, we determine c using the fact that

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X_1,X_2,X_3}(x_1,x_2,x_3)\, dx_3\, dx_2 dx_1$$ should be equal to 1.

When determining the limits of integration, we work again from the outside inwards; x_1 can range from 0 to 1, after which x_2 can range from x_1 to 1, after which x_3 can range from x_2 to 1 (try to do this with another order of integration!):

$$1 = c\int_0^1\int_{x_1}^1\int_{x_2}^1 dx_3\, dx_2 dx_1 = c\int_0^1\int_{x_1}^1 [\int_{x_2}^1 dx_3] dx_2 dx_1$$

$$= c\int_0^1\int_{x_1}^1 (1-x_2) dx_2 dx_1 = c\int_0^1 [\int_{x_1}^1 (1-x_2) dx_2] dx_1$$

$$= c\int_0^1 [x_2 - \tfrac{1}{2}x_2^2]_{x_2=x_1}^1 \, dx_1 = c\int_0^1 \left(\tfrac{1}{2} - x_1 + \tfrac{1}{2}x_1^2\right) dx_1$$

$$= c\int_0^1 \left(\tfrac{1}{2} - x_1 + \tfrac{1}{2}x_1^2\right) dx_1 = c[\tfrac{1}{2}x_1 - \tfrac{1}{2}x_1^2 + \tfrac{1}{6}x_1^3]_{x_1=0}^1 = \frac{1}{6}c \quad \Rightarrow \quad c = 6$$

From the pdf, we can now derive the marginal pdf's:

$$f_{X_1}(x_1) = 6\int_{x_1}^1 \int_{x_2}^1 dx_3 \, dx_2 = 6\int_{x_1}^1 (1 - x_2) dx_2 = 3 - 6x_1 + 3x_1^2 \qquad \text{for } 0 < x_1 < 1$$

$$f_{X_2}(x_2) = 6\int_0^{x_2} \int_{x_2}^1 dx_3 \, dx_1 = 6\int_0^{x_2} (1 - x_2) dx_1 = 6x_2 - 6x_2^2 \qquad \text{for } 0 < x_2 < 1$$

$$f_{X_3}(x_3) = 6\int_0^{x_3} \int_0^{x_2} dx_1 \, dx_2 = 6\int_0^{x_3} x_2 dx_2 = 3x_3^2 \qquad \text{for } 0 < x_3 < 1$$

We can also determine the joint pdf's for any subset of the three random variables. For example, the joint pdf of X_1 and X_2 can be determined by integration over x_3 :

$$f_{X_1,X_2}(x_1, x_2) = 6\int_{x_2}^1 dx_3 = 6(1 - x_2) \qquad \text{for } 0 < x_1 < x_2 < 1$$

This last pdf could have been used as well in finding the marginal pdf's of X_1 and X_2, of course with the same results as above. ◄

5.3 Expected values for functions of two or more random variables

(B&E,pages 172-173)

When we consider a function of the two random variables *X* and *Y,* say *W* = *g*(*X, Y*), it is clear that *W* is an 'ordinary' univariate random variable (not multivariate, because for each outcome of the experiment for which *X* and *Y* have been defined, the outcome *w* of *W* is a scalar, not a vector). We could therefore find the expected value of *W* by first determining the pdf of *W* = *g*(*X, Y*) (see also Chapter 6.1 for techniques to do this) and then use the definition of the expected value for a univariate random variable (see previous course). But we can also do this directly using the joint pdf (given without proof):

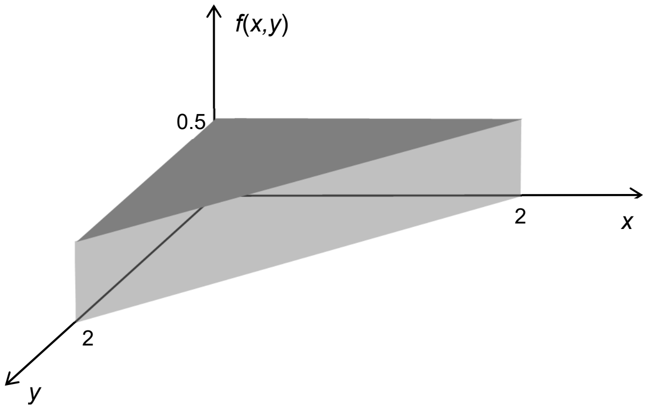
Theorem 5.6 (*B&E, Th.5.2.1*)

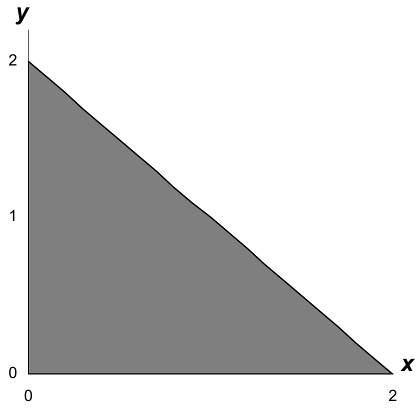
If the pair (*X, Y*) of random variables has the joint pdf $f_{X,Y}(x,y)$, and $W = g(X,Y)$ is an arbitrary real function of *X* and *Y*, then E(*W*) = E(*g*(*X,Y*)), where

$$\mathrm{E}(g(X,Y)) = \begin{cases} \sum_{\text{all } x} \sum_{\text{all } y} g(x,y) \cdot f_{X,Y}(x,y) & \text{if } X \text{ and } Y \text{ discrete} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) \cdot f_{X,Y}(x,y) \, dx \, dy & \text{if } X \text{ and } Y \text{ continuous} \end{cases}$$

Example 5.16

Let the two random variables *X* and *Y* have a joint continuous uniform distribution on the support set with a triangular shape *x* > 0, *y* > 0, *x* + *y* < 2. (Remark: a uniform distribution is a distribution with a constant value for the pdf over the complete support set. For this example, this means that this constant is equal to ½. Check this for yourself using the left figure below!).





To find the correct integration limits, it is more convenient to draw a figure for the support set only (see figure on the right). We can now determine E(XY):

$$\mathrm{E}(XY)=\int_0^2\left(\int_0^{2-x} xy(\tfrac{1}{2})dy\right)dx=$$

$$=\int_0^2\left(\left[\tfrac{1}{4}xy^2\right]_0^{2-x}\right)dx=\int_0^2\left(\left[\tfrac{1}{4}x(2-x)^2\right]\right)dx$$

$$=\int_0^2\left(\tfrac{1}{4}x(4-4x+x^2)\right)dx=$$

$$=\int_0^2(x-x^2+\tfrac{1}{4}x^3)dx$$

$$=\left[\tfrac{1}{2}x^2-\tfrac{1}{3}x^3+\tfrac{1}{16}x^4\right]_0^2=[2-\tfrac{8}{3}+1]=\tfrac{1}{3}$$

Of course we could have found E(XY) as well by changing the order of variables in the integration:

$$\mathrm{E}(XY)=\int_0^2\left(\int_0^{2-y} xy(\tfrac{1}{2})dx\right)dy=\ldots=\frac{1}{3}$$ ◄

Sometimes when we are dealing with double integrals, the amount of work needed to evaluate the integrals can depend on the order of integration. It can be worthwhile to pay attention to this. For example, check that E(XY) for the pdf in Example 5.14 can be easiest determined using the order '*dy dx*' instead of the order '*dx dy*'.

We can apply Theorem 5.6 as well to find the expected value of Y using the joint pdf of X and Y, without explicitly determining first the marginal pdf of Y. We can easily see this by defnining $g(X, Y) = Y$, assuming both X and Y to be continuous:

$$\mathrm{E}(Y)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y\cdot f_{X,Y}(x,y)\,dx\,dy=\int_{-\infty}^{\infty} y\,(\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx)\,dy=\int_{-\infty}^{\infty} y\ f_Y(y)dy=\mathrm{E}(Y)$$

Although both integrals will always result in the same answer, occasionally we might feel the need to indicate whether the expected value was determined by using the pdf of Y or by using the joint pdf, notated by $\mathrm{E}_Y(Y)$ and $\mathrm{E}_{(X,Y)}(Y)$ respectively.

Example 5.17 (Example 5.14 continued)

Given the joint pdf $f_{X,Y}(x,y)=1/20$ for $0<x<10,\ x-1<y<x+1$, we will determine E(Y) now using the joint pdf (with a smartly chosen order of integration!) as well as using the marginal pdf found in Example 5.14.

$$\mathrm{E}(Y) = \mathrm{E}_{(X,Y)}(Y) = \int_0^{10}\int_{x-1}^{x+1} y\frac{1}{20}\,dy\,dx = \int_0^{10}\left[\frac{1}{40}y^2\right]_{y=x-1}^{x+1} dx = \int_0^{10}\frac{1}{10}x\,dx = 5$$

$$\mathrm{E}(Y) = \mathrm{E}_Y(Y) = \int_{-\infty}^{\infty} yf_Y(y)\,dy = \int_{-1}^{1} y\frac{y+1}{20}dy + \int_1^9 y\frac{1}{10}dy + \int_9^{11} y\frac{11-y}{20}dy = \ldots = 5$$

It will be clear that (in this particular case) the second way involves a lot more effort. ◄

Theorem 3.9 (see Prob. Theory and Statistics 1) stated without proof that E(X + Y) = E(X) + E(Y). Now we can formally prove this:

Theorem 5.7 (*B&E, Th.5.2.2*)

If X and Y are random variables, then E(X + Y) = E(X) + E(Y).

Proof

We assume here that X and Y are continuous (otherwise, the integrals should be replaced by summations in the proof below).

$$\begin{aligned}
\mathrm{E}(X+Y) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x+y)\cdot f_{X,Y}(x,y)\,dx\,dy \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x\cdot f_{X,Y}(x,y)\,dx\,dy + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y\cdot f_{X,Y}(x,y)\,dx\,dy \\
&= \int_{-\infty}^{\infty} x\left[\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy\right]dx + \int_{-\infty}^{\infty} y\left[\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx\right]dy \\
&= \int_{-\infty}^{\infty} x\,f_X(x)\,dx + \int_{-\infty}^{\infty} y\,f_Y(y)\,dy = \mathrm{E}_X(X) + \mathrm{E}_Y(Y) = \mathrm{E}(X) + \mathrm{E}(Y)
\end{aligned}$$

We generalise this to k-dimensional random variables.

Theorem 5.8 (*B&E, Th.5.2.1*)

If $\boldsymbol{X} = (X_1,\ldots,X_k)$ has the joint pdf $f_{X_1,\ldots,X_k}(x_1,\ldots,x_k)$ and $Y = g(X_1,\ldots,X_k) = g(\boldsymbol{X})$ is any real function, then $\mathrm{E}(Y) = \mathrm{E}(g(X_1,\ldots,X_k))$, where

$$\mathrm{E}(g(X_1,\ldots,X_k)) = \begin{cases} \displaystyle\sum_{x_1}\cdots\sum_{x_k} g(x_1,\ldots,x_k)\cdot f_{X_1,\ldots,X_k}(x_1,\ldots,x_k) & \text{if } \boldsymbol{X} \text{ discrete} \\ \displaystyle\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} g(x_1,\ldots,x_k)\cdot f_{X_1,\ldots,X_k}(x_1,\ldots,x_k)\,dx_1\ldots dx_k & \text{if } \boldsymbol{X} \text{ continuous} \end{cases}$$

Analogous to Theorem 5.7, we can also show that for any constants $a_1,\ldots,a_k$ we get:

$$\mathrm{E}\left(\sum_{i=1}^{k} a_i X_i\right) = \sum_{i=1}^{k} a_i\,\mathrm{E}(X_i).$$

5.4 Independence of random variables, covariance and correlation

(B&E, pages 149-153, 174-179)

In the previous course, the independence of two *events* A and B has been defined as:

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

When we look again at the table for Example 5.3, we can see that the events $X = 1$ and $Y = 0$ are independent, because $P(X=1,\ Y=0)=0.12$ is equal to $P(X=1)\cdot P(Y=0)=0.4\cdot 0.3=0.12$, or $f_{X,Y}(1,0)=f_X(1)\cdot f_Y(0)$.

		X = Number of units A sold				$f_Y(y)$
		0	1	2	3	
Y = Number of units B sold	0	0.06	0.12	0.08	0.04	0.30
	1	0.06	0.18	0.16	0	0.40
	2	0.20	0.10	0	0	0.30
$f_X(x)$		0.32	0.40	0.24	0.04	1.00

However, the events $X = 0$ and $Y = 0$ are not independent, because $f_{X,Y}(0,0)\neq f_X(0)\cdot f_Y(0)$. So there is a dependency between X and Y, and we say that the random variables X and Y are not independent.

A frequently used definition for the independence of random variables follows below (alternative definitions are also possible, but in the end they are equivalent to each other):

Definition 5.13 (*B&E, Def. 4.4.1*)

Two random variables X and Y are called **independent** (Dutch: onafhankelijk of stochastisch onafhankelijk) if for every $a<b$ and $c<d$ the following equation holds:

$$P(a\le X\le b,\ c\le Y\le d)=P(a\le X\le b)\cdot P(c\le Y\le d)$$

The next theorem follows simply and directly from the definition above:

Theorem 5.9 (*B&E, Th.4.4.1*)

The random variables X and Y are independent if and only if for all x, y:

$$F_{X,Y}(x,y)=F_X(x)\cdot F_Y(y)$$

Furthermore, the random variables X and Y are independent if and only if for all x, y:

$$f_{X,Y}(x,y)=f_X(x)\cdot f_Y(y)$$

Example 5.18

Consider the table for Example 5.1, listing the joint probabilities for the experiment of throwing two dice. For each combination (x,y) we have $f_{X,Y}(x,y)=1/36=1/6\cdot 1/6=f_X(x)\cdot f_Y(y)$. In other words: X and Y are independent. ◄

The theorem makes it easy to find the joint pdf of two random variables if it is known that they are independent, by multiplying the pdf's of X and Y with each other: $f_{X,Y}(x,y)=f_X(x)\cdot f_Y(y)$.

Example 5.19

The table in Example 5.1 could have been found directly by arguing that X and Y are independent, such that we get: $f_{X,Y}(x,y)=f_X(x)\cdot f_Y(y)=1/36$ for $x = 1, \dots, 6$ and $y = 1, \dots, 6$. ◄

The next theorem can be helpful to determine quickly whether two random variables are independent or not. First, we need to introduce the concept **Cartesian product**. The Cartesian product of the sets A and B is notated as $A\times B$ and defined to be $A\times B=\{(x,y)\mid x\in A,\ y\in B\}$.

Theorem 5.10 (*B&E, Th.4.4.2*)

The random variables X and Y with pdf $f_{X,Y}(x,y)$ are independent if and only if:

1. The support of X and Y is a Cartesian product of the supports of X and Y respectively, so $\{(x,y) \mid f_{X,Y}(x,y) > 0\} = \{x \mid f_X(x) > 0\} \times \{y \mid f_Y(y) > 0\}$
2. Some functions $g(\bullet)$ and $h(\bullet)$ exist such that $f_{X,Y}(x,y)$ can be written as $g(x)h(y)$ for all values of x and y belonging to the support of X and Y (we say that the joint pdf can be *factorised*).

Proof

Assume that X and Y are independent. We know that this implies that $f_{X,Y}(x,y) = f_X(x) \cdot f_Y(y)$ holds for all x and y. Consider first values of x and y belonging to the support of X and Y respectively. Then it is directly clear that $f_{X,Y}(x,y) = f_X(x) \cdot f_Y(y)$ can be factorised, so point 2 above will hold. If either x or y does *not* belong to the respective supports, then either $f_X(x) = 0$ or $f_Y(y) = 0$, so $f_{X,Y}(x,y)$ will be 0, meaning that (x, y) does in turn not belong to the support of (X, Y). Thus, point 1 holds as well. You may try yourself to give a proof in the other direction.

Remark. although the functions $g(\bullet)$ and $h(\bullet)$ do not need to be pdf's themselves, they can always be written as $g(x) = c\, f_X(x)$ and $h(x) = 1/c \cdot f_Y(y)$

Example 5.20

Given is the joint pdf $f_{X,Y}(x,y) = 8xy \quad (\text{for } 0 < x < y < 1)$.

This pdf can be factorised (choose for example $g(x) = 8x$ en $h(y) = y$), but X and Y are not independent, because the support set for X and Y $\{(x,y) \mid 0 < x < y < 1\}$ is not a Cartesian product (is not equal to $\{x \mid 0 < x < 1\} \times \{y \mid 0 < y < 1\}$) ◀

Example 5.21

Consider the joint pdf $f_{X,Y}(x,y) = x + y \quad (\text{for } 0 < x < 1,\ 0 < y < 1)$.

Now the support is indeed a Cartesian product, but the pdf can not be factorised, so again X and Y are not independent. ◀

In the previous course, some important results in theorems 3.10, 3.11 and 4.9 were given without proofs. Now we have enough background to do so:

Theorem 5.11 (*B&E, Th.5.2.3, 5.2.6*)

If X and Y are ***independent*** random variables, then:

(1a) $E[g(X) \cdot h(Y)] = E[g(X)] \cdot E[h(Y)]$ for any two real-valued functions $g(\cdot)$ and $h(\cdot)$

(1b) $G_{X+Y}(t) = G_X(t) \cdot G_Y(t)$ (Theorem 3.11, $G(t)$ is the probability generating function)

(1c) $M_{X+Y}(t) = M_X(t) \cdot M_Y(t)$ (Theorem 4.9, $M(t)$ is the moment generating function)

(1d) $E(X \cdot Y) = E(X) \cdot E(Y)$

(2) $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$ (Theorem 3.10)

Proof (for discrete X and Y; replace summations by integrals for continuous random variables)

(1a)
$$E[g(X) \cdot h(Y)] = \sum_x \sum_y g(x)h(y) f_{X,Y}(x,y)$$
$$\overset{\text{indep.}}{=} \sum_x \sum_y g(x)h(y) f_X(x) f_Y(y) = \sum_x g(x)\, f_X(x) \sum_y h(y)\, f_Y(y)$$
$$= E(g(X)) \cdot E(h(Y))$$

(1b) follows from 1a :

$$G_{X+Y}(t) = \mathrm{E}[t^{X+Y}] = \mathrm{E}[t^X t^Y] \overset{(1a)}{=} \mathrm{E}[t^X]\mathrm{E}[t^Y] = G_X(t)G_Y(t)$$

(1c) follows from 1a :

$$M_{X+Y}(t) = \mathrm{E}[e^{t(X+Y)}] = \mathrm{E}[e^{tX}e^{tY}] \overset{(1a)}{=} \mathrm{E}[e^{tX}]\mathrm{E}[e^{tY}] = M_X(t)M_Y(t)$$

(1d) follows directly from 1a

(2) $\mathrm{Var}(X+Y) = \mathrm{E}((X+Y)^2) - (\mathrm{E}(X+Y))^2 =$ (Theorem 5.7)

$= \mathrm{E}(X^2 + Y^2 + 2XY) - (\mathrm{E}(X) + \mathrm{E}(Y))^2 =$

$= \mathrm{E}(X^2) + \mathrm{E}(Y^2) + 2\mathrm{E}(XY) - ([\mathrm{E}(X)]^2 + [\mathrm{E}(Y)]^2 + 2\,\mathrm{E}(X)\mathrm{E}(Y)) =$

$= \mathrm{E}(X^2) - [\mathrm{E}(X)]^2 + \mathrm{E}(Y^2) - [\mathrm{E}(Y)]^2 + 2\mathrm{E}(XY) - 2\,\mathrm{E}(X)(\mathrm{E}Y) =$ (use (1d))

$= \mathrm{E}(X^2) - [\mathrm{E}(X)]^2 + \mathrm{E}(Y^2) - [\mathrm{E}(Y)]^2 = \mathrm{Var}(X) + \mathrm{Var}(Y)$

Note that we used (1d) in the proof for Theorem 5.11(2) . So, when X and Y are not independent, we can NOT conclude that $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$.
But from the third line in the proof of (2), we can see that ALWAYS:

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,[\,\mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y)\,].$$

The term $\mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y)$ will be called the *covariance*:

Definition 5.14 (*B&E, Def. 5.2.1*)

The **covariance** of two random variables X and Y, (notation σ_{XY} or $\mathrm{Cov}(X,Y)$, is defined as:

$$\mathrm{Cov}(X,Y) = \mathrm{E}[(X - \mathrm{E}(X))(Y - \mathrm{E}(Y))] = \mathrm{E}[(X - \mu_X)(Y - \mu_Y)]$$

Theorem 5.12 (*B&E, Th.5.2.5*)

The covariance of two random variables X and Y can also be written as:

$$\mathrm{Cov}(X,Y) = \mathrm{E}(XY) - \mu_X\mu_Y$$

Furthermore, $\mathrm{Cov}(X, Y) = 0$ if X and Y are independent.

Proof

$$\mathrm{Cov}(X,Y) = \mathrm{E}[(X-\mu_X)(Y-\mu_Y)] = \mathrm{E}[XY - \mu_X Y - X\mu_Y + \mu_X\mu_Y]$$

$$= \mathrm{E}(XY) - \mathrm{E}(\mu_X Y) - \mathrm{E}(X\mu_Y) + \mathrm{E}(\mu_X\mu_Y) = \mathrm{E}(XY) - \mu_X\mathrm{E}(Y) - \mu_Y\mathrm{E}(X) + \mu_X\mu_Y$$

$$= \mathrm{E}(XY) - \mu_X\mu_Y$$

Applying Theorem 5.11 (1d) gives the result that the covariance is 0 if X and Y independent.

Example 5.22

We again consider the random variables X and Y from Example 5.2:

		X = Number of units A sold				
		0	1	2	3	$f_Y(y)$
Y = Number of units B sold	0	0.06	0.12	0.08	0.04	0.30
	1	0.06	0.18	0.16	0	0.40
	2	0.20	0.10	0	0	0.30
$f_X(x)$		0.32	0.40	0.24	0.04	1.00

This example has be chosen such that E(X) and E(Y) are both exactly 1, as can be checked easily using the marginal pdf's. Below, we use both formulas to find the covariance:

(1) $\text{Cov}(X,Y) = \text{E}(XY) - \mu_X\mu_Y = \sum_{x=0}^{3}\sum_{y=0}^{2} x\,y\, f_{X,Y}(x,y) - 1\times 1$

$= 1\times1\times0.18 + 1\times2\times0.10 + 2\times1\times0.16 - 1 = -0.30$

(2) $\text{Cov}(X,Y) = \text{E}[(X-\mu_X)(Y-\mu_Y)] = \text{E}[(X-1)(Y-1)] = \sum_{y=0}^{2}\sum_{x=0}^{3}(x-1)(y-1) f_{X,Y}(x,y) =$

$$\begin{aligned}
&= + (0-1)\times(0-1)\times0.06 + (1-1)\times(0-1)\times0.12 + (2-1)\times(0-1)\times 0.08 + (3-1)\times(0-1)\times 0.04 + \\
&\quad + (0-1)\times(1-1)\times0.06 + (1-1)\times(1-1)\times0.18 + (2-1)\times(1-1)\times 0.16 + (3-1)\times(1-1)\times 0.00 + \\
&\quad + (0-1)\times(2-1)\times0.20 + (1-1)\times(2-1)\times0.10 + (2-1)\times(2-1)\times 0.00 + (3-1)\times(2-1)\times 0.00 = \\
&= 0.06 - 0.08 - 0.08 - 0.20 = -0.30
\end{aligned}$$

Note that applying the first formula is a lot simpler here; this will be so in most cases. ◄

Example 5.23 (Example 5.16 continued).

Consider again the joint uniform distribution on the triangular support $x > 0$, $y > 0$, $x + y < 2$. The expectations E(X) and E(Y) are both 2/3 (check (!), e.g. by first finding the marginal pdf's, or use Theorem 5.6). E($X\,Y$) has been found already in Example 5.16 (=1/3). The covariance is:

$$\text{Cov}(X,Y) = \text{E}(XY) - \text{E}(X)\text{E}(Y) = \tfrac{1}{3} - (\tfrac{2}{3})(\tfrac{2}{3}) = -\tfrac{1}{9}$$ ◄

The covariance tells us something about the linear relationship between two random variables. We can see this when we take a look at the definition of the covariance in the discrete case (similar argument in the continuous case):

$$\text{Cov}(X,Y) = \text{E}[(X-\mu_X)(Y-\mu_Y)] = \sum_{\text{all } x}\sum_{\text{all } y}(x-\mu_X)(y-\mu_Y)\cdot f_{X,Y}(x,y)$$

When relatively large values for X (values larger than E(X)) are linked to relatively large values for Y, (and vice-versa), then the product $(x-\mu_X)(y-\mu_Y)$ in the above formula will be positive. Since the covariance is a weighted average of these products, it will be positive in case values for X and Y tend to move in the same direction. If, however, relatively large values for X are more often linked to relatively small values for Y, (and vice-versa), then this will result in a negative covariance.

Theorem 5.13 (*B&E, Th.5.2.4*)

For any random variables X, Y and U and constants a and b , we have:

Cov(aX , bY) = ab Cov(X , Y)

Cov($X + a$, $Y + b$) = Cov(X , Y)

Cov(X , $aX + b$) = a Var(X)

Cov($X + U$, Y) = Cov(X , Y) + Cov(U , Y)

Proof

See problem 5.29.

Theorem 5.14 (*≈B&E, Th.5.2.6*)

If X and Y are any two random variables, then:

$$\text{Var}(aX + bY + c) = a^2\,\text{Var}(X) + b^2\,\text{Var}(Y) + 2ab\text{Cov}(X, Y)$$

Proof

Do this yourself, see proof for part (2) of Theorem 5.11.

Example 5.24 (continuation of Example 5.22)

Define $V = X + Y$ for X and Y in Example 5.22. The pdf for V is 0.06, 0.18, 0.46 and 0.30 for $v =$ 0, 1, 2 and 3 respectively (check). Using the definition for the variance, this results in Var(V) = 0.72.

But it is also possible to find the variance of $V = X + Y$ by applying Theorem 5.14:

$$\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y) = 0.72 + 0.6 + 2\ (-0.30) = 0.72$$

(Using the table in Example 5.22, we can find Var(X) = 0.72 and Var(Y) = 0.6)

Similarly, the variance of $X - Y$ (note the plus sign for the term Var(Y) in the next formula; this follows directly from Theorem 5.14 with $a = 1$ and $b = -1$!!!):

$$\text{Var}(X-Y) = \text{Var}(X) + \text{Var}(Y) - 2\text{Cov}(X, Y) = 0.72 + 0.6 - 2\ (-0.30) = 1.92$$ ◄

The value for the covariance depends on the chosen measurement scale (for example, mm or cm, kg or g) of the variables. Thus, when a distance X is measured in meters, the covariance will be a hundred times as small as the covariance when X would be measured in centimeters. This makes the value of the covariance difficult to interpret; the size of the covariance says nothing about the *strength* of the linear relationship between two random variables. However, when the covariance is divided by the standard deviations of both variables, we arrive at the correlation coefficient, which can be seen as a kind of standardised covariance:

Definition 5.15 (*B&E, Def. 5.3.1*)

The **coefficent of correlation (Dutch: correlatiecoëfficiënt)** of two random variables X and Y is defined by:

$$\rho_{X,Y} = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y}$$

(only when σ_X and σ_Y both greater than 0).

Theorem 5.15 (*B&E, Th.5.3.1*)

If $\rho_{X,Y}$ is the coefficient of correlation of two random variables X and Y, then:

$$-1 \le \rho_{X,Y} \le 1$$

and $\rho_{X,Y}$ is either 1 or -1 if and only if some $a \neq 0$ and b exist such that $P(\ Y = aX + b\ ) = 1$.

Proof

Define $W = \dfrac{Y}{\sigma_Y} - \rho_{X,Y}\dfrac{X}{\sigma_X}$. (Many will, understandably, wonder why W is defined in this particular way. The only reason: the desired proof becomes rather simple this way….)

Using Theorem 5.14 (recall that $\sigma_X^2 = \text{Var}(X)$ and $\sigma_Y^2 = \text{Var}(Y)$):

$$\begin{aligned}\text{Var}(W) &= \frac{1}{\sigma_Y^2}\sigma_Y^2 + \frac{\rho_{X,Y}^2}{\sigma_X^2}\sigma_X^2 - 2\rho_{X,Y}\frac{\text{Cov}(X,Y)}{\sigma_X\sigma_Y} \\ &= 1 + \rho_{X,Y}^2 - 2\rho_{X,Y}^2 \\ &= 1 - \rho_{X,Y}^2\end{aligned}$$

But since any variance, including $\text{Var}(W)$, is never negative, we get:

$$1 - \rho_{X,Y}^2 \ge 0 \quad \Rightarrow \quad -1 \le \rho_{X,Y} \le 1$$

which proves the first part of the theorem.

To prove the second part, we first note that whenever $\rho_{X,Y} = \pm 1$, it follows from the above that $\mathrm{Var}(W) = 0$. This must mean that W can only attain one single value, and thus it follows that $W = \dfrac{Y}{\sigma_Y} - \rho_{X,Y}\dfrac{X}{\sigma_X}$ defines a linear relation between X and Y such that $Y = aX + b$ for some a and b (the values for a and b are of no interest to us, the only important thing is the conclusion that X and Y are linearly related).

Conversely, if $Y = aX + b$, then it follows from Theorem 5.14 that $\mathrm{Var}(Y) = a^2\,\mathrm{Var}(X)$, so $\sigma_Y = \sqrt{a^2}\sigma_X = |a|\sigma_X$. Also:

$$\mathrm{Cov}(X, Y) = \mathrm{Cov}(X, aX + b) = a\mathrm{Cov}(X, X) = a\mathrm{Var}(X)$$

such that

$$\rho_{X,Y} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X\sigma_Y} = \frac{a\sigma_X^2}{|a|\sigma_X^2} = \begin{cases} 1 & \text{if } a > 0 \\ -1 & \text{if } a < 0 \end{cases}$$

So when a (perfect) linear relation exists between X and Y, the coefficient of correlation is either +1 or −1; if the relation is almost perfect, the coefficient of correlation will be close to either + 1 or −1.

Example 5.25 *(Continuation of Example 5.14)*

Consider the joint pdf $f_{X,Y}(x, y) = 1/20 \quad \text{for } 0 < x < 10,\ x-1 < y < x+1$.

Simply by looking at the figure for the support set, we expect the coefficient of correlation to be close to 1. This is indeed the case, as follows from the calculations below:

$$\mathrm{E}(X) = \int_0^{10}\int_{x-1}^{x+1} x\frac{1}{20}\,dy\,dx = \int_0^{10} x\frac{2}{20}\,dx = 5$$

$$\mathrm{E}(X^2) = \int_0^{10}\int_{x-1}^{x+1} x^2\frac{1}{20}\,dy\,dx = \int_0^{10} x^2\frac{1}{10}\,dx = \frac{100}{3} \quad\Rightarrow\quad \mathrm{Var}(X) = \frac{100}{3} - 5^2 = \frac{25}{3}$$

$$\mathrm{E}(Y) = \int_0^{10}\int_{x-1}^{x+1} y\frac{1}{20}\,dy\,dx = \int_0^{10} x\frac{4}{40}\,dx = 5$$

$$\mathrm{E}(Y^2) = \int_0^{10}\int_{x-1}^{x+1} y^2\frac{1}{20}\,dy\,dx = \frac{1}{60}\int_0^{10}[(x+1)^3 - (x-1)^3]\,dx = \frac{1}{240}[(x+1)^4 - (x-1)^4]_0^{10} = \frac{101}{3} \quad\Rightarrow\quad \mathrm{Var}(Y) = \frac{26}{3}$$

$$\mathrm{E}(XY) = \int_0^{10}\int_{x-1}^{x+1} xy\frac{1}{20}\,dy\,dx = \frac{1}{40}\int_0^{10} 4x^2\,dx = \frac{100}{3} \quad\Rightarrow\quad \mathrm{Cov}(X,Y) = \frac{100}{3} - 5\cdot 5 = \frac{25}{3}$$

$$\rho_{X,Y} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X\sigma_Y} = \frac{25/3}{\sqrt{(25/3)(26/3)}} = \frac{5}{\sqrt{26}} \approx 0.981$$ ◄

When the coefficient of correlation of X and Y is equal to 0, we say that X and Y are *uncorrelated*. Whenever X and Y are independent, it is easy to see that the covariance, and thus the coefficient of correlation, is equal to 0. In other words: if X and Y are independent, then they are uncorrelated as well. However, the reverse is not always true, as we can see in the next example.

Example 5.26

Consider the joint probability table below:

		X			
		-1	0	1	
	-1	0.1	0	0.1	0.2
Y	0	0	0.6	0	0.6
	1	0.1	0	0.1	0.2
		0.2	0.6	0.2	

Using the marginal pdf's, it is very easy to see that E(*X*) = E(*Y*) = 0.
Furthermore, we get: E(*XY*) = (-1)(-1)0.1 + (1)(-1)0.1 + (-1)(1)0.1 + (1)(1)0.1 = 0.
This results in the covariance: Cov(*X*, *Y*) = 0 – 0 = 0, so the two random variables are uncorrelated. However, it is also clear that they are not independent, since $f_{X,Y}(x,y) \neq f_X(x) \cdot f_Y(y)$. ◄

Example 5.27

Let $X \sim N(0, 1)$, and define $Y = X^2$. Obviously, *X* and *Y* are not independent, since knowing the value for *X* determines completely the value of *Y*. Still, we get:

$$\mathrm{Cov}(X,Y) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y) = \mathrm{E}(XX^2) - \mathrm{E}(X)\mathrm{E}(X^2) =$$

$$= \mathrm{E}(X^3) - 0 \cdot 1 = \mathrm{E}(X^3) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^3 e^{-\frac{1}{2}x^2} dx$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0} x^3 e^{-\frac{1}{2}x^2} dx + \frac{1}{\sqrt{2\pi}} \int_{0}^{\infty} x^3 e^{-\frac{1}{2}x^2} dx = 0$$

(note that the last two terms cancel each other). ◄

Multivariate

We generalise the results above to *k*-dimensional random variables.

Definition 5.16 (*B&E, Def. 4.4.1*)

The random variables $(X_1,...,X_k)$ are called **mutually (or stochastically) independent** (Dutch: onderling independent, stochastisch independent) if for every $a_i < b_i$ we get:

$$P(a_1 \leq X_1 \leq b_1, \dots,\ a_k \leq X_k \leq b_k) = P(a_1 \leq X_1 \leq b_1) \times\ \dots\ \times P(a_k \leq X_k \leq b_k)$$

Remark. There are situations where $(X_1,...,X_k)$ are not mutually independent, but where nevertheless all X_i are independent of all X_j for all i and j such that i ≠ j (see for example Problem 5.48). When this is the case, we say that $(X_1,...,X_k)$ are pairwise independent; note that this does not imply mutual independence (unless k = 2, because then there is no difference between the two concepts of independence). The book of Bain and Engelhardt simply talks about 'independence', where they actually mean 'mutual independence' (and not 'pairwise independence'). In the remainder, we will often just use the term 'independence', even when actually 'mutual independence' is meant.

Theorem 5.16 (*B&E, Th.4.4.1*)

The random variables $X_1,...,X_k$ are mutually independent if and only if for all $x_1,...,x_k$ we get

$$F_{(X_1,...,X_k)}(x_1,...,x_k) = F_{X_1}(x_1) \cdots F_{X_k}(x_k)$$

Or if and only if:

$$f_{(X_1,...,X_k)}(x_1,...,x_k) = f_{X_1}(x_1) \cdots f_{X_k}(x_k)$$

Example 5.28

The random variables X_1, X_2 and X_3 in Example 5.15 are dependent, since

$$f_{X_1,X_2,X_3}(x_1,x_2,x_3) \neq f_{X_1}(x_1) f_{X_2}(x_2) f_{X_3}(x_3).$$ ◄

Theorem 5.11 can be generalised as well in a straightforward way. We do that below only for the most important result:

Theorem 5.17 *(B&E, Th.6.4.1)*

If the random variables $X_1,...,X_k$ are mutually independent, then:

$$M_{X_1+...+X_k}(t) = \prod_{i=1}^{k} M_{X_i}(t)$$

Nothing is changed with respect to the definitions for the covariance and the coefficient of correlation, because they are only defined for pairs of random variables. We sometimes use the abbreviated notations: σ_i^2 instead of $\sigma_{X_i}^2$ and σ_{ij} instead of $\sigma_{X_iX_j}$.

Repeated application of Theorem 5.14 and Theorem 5.13 results in the following (see problem 5.38).

Theorem 5.18 *(B&E, Eq.5.2.14)*

If $X_1,...,X_k$ are random variables , then :

$$\mathrm{Var}(\sum_{i=1}^{k} a_i X_i) = \sum_{i=1}^{k} a_i^2 \mathrm{Var}(X_i) + 2\sum\sum_{i<j} a_i a_j \mathrm{Cov}(X_i, X_j) = \sum_{i=1}^{k} a_i^2 \sigma_i^2 + 2\sum\sum_{i<j} a_i a_j \sigma_{ij}$$

For the ***independent*** random variables $X_1,...,X_k$ we get (note: pairwise independence is sufficient here):

$$\mathrm{Var}(\sum_{i=1}^{k} a_i X_i) = \sum_{i=1}^{k} a_i^2 \mathrm{Var}(X_i) = \sum_{i=1}^{k} a_i^2 \sigma_i^2$$

$\mathrm{var}(\sum_{i=1}^{k-1} a_i x_i + a_k x_k) = 2\mathrm{cov}(\sum_{i=1}^{k-1} a_i x_i, a_k x_k) = 2\sum_{i=1}^{k-1} a_i a_k \sigma_{ik}$.

Example 5.29

Suppose $Y \sim \mathrm{Bin}(n, p)$. We saw earlier that Y can be seen as the sum of n mutually independent Bernoulli random variables X_i. For each i it follows that: $\mathrm{E}(X_i) = p$, and $\mathrm{Var}(X_i) = pq$.

Then: $\mathrm{E}(Y) = \mathrm{E}(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} \mathrm{E}(X_i) = np$ and $\mathrm{Var}(Y) = \mathrm{Var}(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} \mathrm{Var}(X_i) = npq$. ◄

As illustrated in the theorem above, we often need the covariances between each combination of two random variables from the set $X_1,...,X_k$. It can be convenient to group those in a **variance-covariance matrix**, often indicated by the Greek capital Σ (sigma):

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1k} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{k1} & \sigma_{k2} & \cdots & \sigma_k^2 \end{pmatrix}$$

Note that this matrix is symmetrical, since $\sigma_{ij} = \sigma_{ji}$.

Using matrix notation, we can write $\mathrm{Var}(\sum_{i=1}^{k} a_i X_i)$ as:

$\mathrm{Var}(\boldsymbol{a}\boldsymbol{X}^T) = \boldsymbol{a}\Sigma\boldsymbol{a}^T$, where $\boldsymbol{a} = (a_1, a_2, ..., a_k)$ and $\boldsymbol{X} = (X_1, X_2, ..., X_k)$

Remark. In the literature, often vertical vectors $\boldsymbol{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_k \end{pmatrix}$ are being used, instead of the horizontal vectors as used by Bain and Engelhardt. This would result in $\mathrm{Var}(\boldsymbol{a}^T\boldsymbol{X}) = \boldsymbol{a}^T\Sigma\boldsymbol{a}$.

5.5 Conditional distributions

(B&E, pages 153-158, 180-183)

Suppose we have for all households in a certain population data on the number of children and the number of rooms in the house in which the household resides. The question: "What is the distribution of the number of rooms in households with three children?" is a question about a *conditional distribution*. The same applies to the question: 'what is the distribution of the number of children in households with five rooms?"

To answer the first question, we extract from the population all households with three children and we determine the distribution of the number of rooms in the house in this subpopulation. It is very likely that this distribution will be different from the *unconditional* distribution of the number of rooms (which is the distribution of the number of rooms taken over all households in the population).

Example 5.30

In Example 5.2, the joint pdf was given for the random variables *X* and *Y,* representing the number of units sold of A and B respectively. Suppose we are interested in the distribution of *Y*, given that *X* = 1. We will first determine $P(Y=0\,|\,X=1)$, using Definition 2.1 of the previous course:

$$P(Y=0\,|\,X=1)=\frac{P(Y=0,\ X=1)}{P(X=1)}=\frac{f_{X,Y}(1,\ 0)}{f_X(1)}=\frac{0.12}{0.4}=0.3$$

In a similar manner, we find: $P(Y=1\,|\,X=1)=0.45$ and $P(Y=2\,|\,X=1)=0.25$. Note that these three probabilities together represent a complete pdf (non-negative probabilities, sum equals 1). This conditional pdf can be written as $f_{Y|X}(y\,|\,1)$, with values $f_{Y|X}(0\,|\,1)=0.3$, $f_{Y|X}(1\,|\,1)=0.45$ and $f_{Y|X}(2\,|\,1)=0.25$. So $f_{Y|X}(y\,|\,1)$ is the pdf of *Y,* given that *X* = 1.

For *X* = 2, we will have another pdf, i.e. $f_{Y|X}(y\,|\,2)$, with values $f_{Y|X}(0\,|\,2)=1/3$ and $f_{Y|X}(1\,|\,2)=2/3$. We can say that $f_{Y|X}(y\,|\,x)$ represents a whole family of pdf's, one for each possible value of *x*. ◄

Definition 5.17 (*B&E, Def. 4.5.1*)

If (*X*, *Y*) has the joint pdf $f_{X,Y}$ (*x*, *y*), then the **conditional** (Dutch: voorwaardelijke of conditionele) **pdf** of *Y* given *X* = *x* is defined as:

$$f_{Y|X}(y\,|\,x)=\frac{f_{X,Y}(x,y)}{f_X(x)} \qquad \text{(alternative notation: } f_{Y|X=x}(y)\text{)}$$

for values of *x* with $f_X(x)>0$.

So the conditional pdf is the ratio of the joint pdf and the marginal pdf. Of course, we can only meaningfully consider the conditional distribution of *Y* given *X* = *x* whenever $f_X(x)>0$. Strictly speaking, $f_{Y|X}(y\,|\,x)$ is therefore not defined whenever $f_X(x)=0$. But since this should be obvious, this will not be stated explicitly each time.

Remark. As mentioned in the example above, Definition 2.1 of the course Prob. Theory and Statistics 1 results directly in Definition 5.17 for *discrete* random variables. For *continuous* random variables, we cannot apply Definition 2.1 directly (because *P*(*X* = *x*) = 0), but nevertheless the same definition remains useful. To see this, we apply Definition 2.1 (for ε > 0 and δ > 0):

$$P(y<Y\le y+\varepsilon\,|\,x<X\le x+\delta)=\frac{P(x<X\le x+\delta\cap y<Y\le y+\varepsilon)}{P(x<X\le x+\delta)}$$

For small values of δ and ε, this gives approximately the result

$$\varepsilon\, f_{Y|X=x}(y\,|\,x)=\frac{\delta\times\varepsilon\times f_{X,Y}(x,y)}{\delta\times f_X(x)}$$

which is the formula Definition 5.17.

Remark. Instead of the unambiguous notation in Definition 5.17, other books often use the notation f(y | x). While this notation is has the advantage of being shorter, it is actually quite ambiguous, and therefore not recommended.

Of course, by interchanging *X* with *Y*, it follows that: $f_{X|Y}(x\,|\,y) = \dfrac{f_{X,Y}(x,y)}{f_Y(y)}$.

Note that $f_{Y|X}(y\,|\,x)$ can be considered to be a whole family of pdf's for *Y*; each possible value of *x*.defines another pdf. (We could say that *x* is now a parameter of the conditional distribution!). The usual properties for a pdf should remain true, i.e.:

$$f_{Y|X}(y\,|\,x) \geq 0 \quad \text{(for all } y\text{)} \quad \text{and}$$

$$\int_{-\infty}^{\infty} f_{Y|X}(y\,|\,x)\,dy = 1 \ \ \text{(if } Y \text{ continuous)} \quad \text{or} \quad \sum_{\text{all } y} f_{Y|X}(y\,|\,x) = 1 \quad \text{(if } Y \text{ discrete)}$$

The first property follows directly from Definition 5.17. For the second property (assume *Y* to be continuous; in the discrete case we replace the integrals by summations):

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

The conditional CDF (of *Y* given *X* = *x*) follows then directly:

$$F_{Y|X}(y\,|\,x) = P(Y \leq y\,|\,X = x) = \int_{-\infty}^{y} f_{Y|X}(t\,|\,x)\,dt$$

Example 5.31

Assume that $(X,Y) \sim \text{MULT}(n, p_1, p_2)$. We already know (see page 9) that $X \sim \text{BIN}(n, p_1)$, $Y \sim \text{BIN}(n, p_2)$ and $f_{X,Y}(x,y) = \dfrac{n!}{x!\,y!\,(n-x-y)!} p_1^{\,x} p_2^{\,y} (1-p_1-p_2)^{(n-x-y)}$.

The conditional pdf of *Y* given *X* = *x* is determined by:

$$f_{Y|X}(y\,|\,x) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{\dfrac{n!}{x!\,y!\,(n-x-y)!} p_1^{\,x} p_2^{\,y} (1-p_1-p_2)^{n-x-y}}{\dfrac{n!}{x!\,(n-x)!} p_1^{\,x} (1-p_1)^{n-x}}$$

$$= \frac{\dfrac{1}{y!\,(n-x-y)!} p_2^{\,y} (1-p_1-p_2)^{n-x-y}}{\dfrac{1}{(n-x)!} (1-p_1)^{y+n-x-y}}$$

$$= \frac{(n-x)!}{y!\,(n-x-y)!} \left(\frac{p_2}{1-p_1}\right)^{y} \left(\frac{1-p_1-p_2}{1-p_1}\right)^{n-x-y}$$

$$= \binom{n-x}{y} \left(\frac{p_2}{1-p_1}\right)^{y} \left(1 - \frac{p_2}{1-p_1}\right)^{(n-x-y)}$$

Since this is the pdf for a binomial distribution, we know now that $Y\,|\,X = x \sim \text{BIN}(n-x, \dfrac{p_2}{1-p_1})$.

Check for yourself that this is a logical result! (Hint: given *X* = *x*, the remaining *n* – *x* trials can only assume outcomes in either the second or the third category, with 'new' probabilities which should again add up to 1.) ◀

Example 5.32 (Example 5.13 continued)

Let $f_{X,Y}(x,y) = 2e^{-x-y}$ for $0 \le x \le y$. Earlier, we determined the marginal densities: $f_X(x) = 2e^{-2x}$ for $x \ge 0$, and $f_Y(y) = 2e^{-y}(1-e^{-y})$ for $y \ge 0$.

The conditional pdf of Y given $X = x$, with $x \ge 0$, is:

$$f_{Y|X}(y\,|\,x) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{2e^{-x-y}}{2e^{-2x}} = e^{x-y} \quad \text{for } y \ge x \ (\ge 0).$$

Note that this is the pdf of a two-parameter exponential distribution (Definition 4.8), so:

$$Y \,|\, X = x \sim \text{EXP}(1, x).$$

The conditional CDF of Y given $X = x$ is:

$$F_{Y|X}(y\,|\,x) = \int_{-\infty}^{y} f_{Y|X}(t\,|\,x)dt = \int_{x}^{y} e^{x-t}dt = \left[-e^{x-t}\right]_{t=x}^{t=y} = 1 - e^{x-y} \quad \text{for } y \ge x \ (\ge 0)$$

(For $y < x$: $F_{Y|X}(y\,|\,x) = 0$)

And the conditional pdf of X given $Y = y$, with $y \ge 0$, is:

$$f_{X|Y}(x\,|\,y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{2e^{-x-y}}{2e^{-y}(1-e^{-y})} = \frac{e^{-x}}{1-e^{-y}} \quad \text{for } 0 \le x \le y.$$

Note that this is not one of the distributions we discussed before (it is referred to as a 'truncated exponential distribution'). ◄

Theorem 5.19 (≈*B&E, Th.4.5.1*)

If the random variables X and Y are independent, then:

$$f_{Y|X}(y\,|\,x) = f_Y(y) \quad \text{(for all } y\text{, and for all } x \text{ with } f_X(x) > 0\text{)}$$

and:

$$f_{X|Y}(x\,|\,y) = f_X(x) \quad \text{(for all } x\text{, and for all } y \text{ with } f_Y(y) > 0\text{)}$$

Proof

Use Definition 5.17: $f_{Y|X}(y\,|\,x) = \dfrac{f_{X,Y}(x,y)}{f_X(x)} \overset{\text{indep.}}{=} \dfrac{f_X(x)f_Y(y)}{f_X(x)} = f_Y(y)$

In words, this theorem states that for independent random variables X and Y, knowledge of the value the random variable X assumes does not result in another distribution of Y, and vice versa. Try to see the logic! To show that X and Y are not independent, it is sufficient to show that $f_{Y|X}(y\,|\,x)$ really depends on x (or that $f_{X|Y}(x\,|\,y)$ depends on y). For example, the conditional pdf's in Example 5.32 show immediately that X and Y are not independent.

In order to find the joint pdf of X and Y, it can be helpful to see if the conditional pdf of Y given $X=x$ (or the other way around) can be found easily. In those situations, the following theorem, which follows directly from Definition 5.17, is very useful:

Theorem 5.20 (*B&E, Th.4.5.1*)

If X and Y are random variables, then:

$$f_{X,Y}(x,y) = f_X(x)f_{Y|X}(y\,|\,x) = f_Y(y)f_{X|Y}(x\,|\,y)$$

Example 5.33

A fair die is thrown and results in a random score X (=number of dots). Afterwards, a fair coin will be tossed X times. When we define Y as the total number of times that the coin shows 'Head', then it is easy to see that the conditional distribution of Y given $X = x$ is binomial, so $Y \mid X=x \sim \text{Bin}(x\ ,\ ½)$. The joint pdf is shown in the table below, by using the fact that $f_{X,Y}(x,y) = f_X(x) f_{Y|X}(y \mid x)$.

	$y = 0$	$y = 1$	$y = 2$	$y = 3$	$y = 4$	$y = 5$	$y = 6$	$f_X(x)$
$x = 1$	1/6×1/2	1/6×1/2	0	0	0	0	0	1/6
$x = 2$	1/6×1/4	1/6×2/4	1/6×1/4	0	0	0	0	1/6
$x = 3$	1/6×1/8	1/6×3/8	1/6×3/8	1/6×1/8	0	0	0	1/6
$x = 4$	1/6×1/16	1/6×4/16	1/6×6/16	1/6×4/16	1/6×1/16	0	0	1/6
$x = 5$	1/6×1/32	1/6×5/32	1/6×10/32	1/6×10/32	1/6×5/32	1/6×1/32	0	1/6
$x = 6$	1/6×1/64	1/6×6/64	1/6×15/64	1/6×20/64	1/6×15/64	1/6×6/64	1/6×1/64	1/6
$f_Y(y)$	63/384	120/384	99/384	64/384	29/384	8/384	1/384	

Or, in formula: $f_{X,Y}(x,y) = \frac{1}{6}\binom{x}{y}(0.5)^x$. Note that the last row shows the marginal pdf of Y by summing over all possible values for x. ◄

Example 5.34

Let X be a uniform distribution on [1 , 3] (so with density = ½). Assume that the conditional distribution of Y given $X = x$ is uniform on the interval $[2x - 1\ ,\ 4x-1]$, which means that the conditional pdf is:

$$f_{Y|X}(y \mid x) = \frac{1}{(4x-1)-(2x-1)} = \frac{1}{2x} \quad \text{for } 2x - 1 \le y \le 4x - 1 \text{ (only defined when } 1 \le x \le 3\text{)}.$$

The joint distribution has the density function:

$$f_{X,Y}(x,y) = f_X(x)\ f_{Y|X}(y \mid x) = \frac{1}{2}\left(\frac{1}{2x}\right) = \frac{1}{4x} \quad \text{on the set } \{(x,y) \mid 1 \le x \le 3\ ;\ 2x - 1 \le y \le 4x - 1\}$$ ◄

We have seen earlier that, for discrete random variables, the probability of an event A (defined on the same sample space as X and Y) can be written as: $P(A) = P((X,Y) \in A) = \sum\sum_{(x,y)\in A} f_{X,Y}(x,y)$. Because of Theorem 5.20, we can find the probability of A as well by using the conditional pdf of Y given $X = x$ and the pdf of X:

$$\begin{aligned} P(A) = P((X,Y) \in A) &= \sum\sum_{(x,y)\in A} f_{X,Y}(x,y) \\ &= \sum\sum_{(x,y)\in A} f_{Y|X}(y \mid x) f_X(x) \\ &= \sum_x \Big[\sum_{y|(x,y)\in A} f_{Y|X}(y \mid x) \Big] f_X(x) \\ &= \sum_x P(A \mid X = x) f_X(x) \end{aligned}$$

Note that this result is just an application of the Law of Total Probability (Theorem 2.8, Prob. Theory and Statistics 1).

For continuous random variables, we get an analogous result:

$$P(A) = \int_{-\infty}^{\infty} P(A \mid X = x) f_X(x)\, dx$$

When we use the above formula, we say that we find the probability of A by *conditioning on X.* This can be a very helpful when $P(\ A \mid X = x\ )$ can be determined more easily than $P(\ A\ )$ itself.

Example 5.35 (Example 5.33 continued)

In the experiment as described in Example 5.33, it is not immediately clear how we can find the (marginal) probability $P(Y = 3)$. But by conditioning on X, we obtain:

$$P(Y=3)=\sum_x P(Y=3\,|\,X=x)f_X(x)=1/6\cdot(1/8+4/16+10/32+20/64)=64/384 \qquad ◄$$

Example 5.36 (Example 5.34 continnued)

For the pdf in Example 5.34, we will determine the probability that Y is at least twice as large as X. We could find this probability by evaluating a double integral over the joint pdf on the region where $y > 2x$. Here, we will do this by conditioning on X (note the subtle but important difference between the second and third expression below!):

$$P(Y>2X)=\int_1^3 P(Y>2X\,|\,X=x)f_X(x)dx=\int_1^3 P(Y>2x\,|\,X=x)f_X(x)dx=$$

Because the conditional distribution of Y given $X = x$ is the uniform distribution on the interval $[2x-1\ ,\ 4x-1]$, it follows directly that (check!!)

$$P(Y>2x\,|\,X=x)=\frac{(4x-1)-2x}{2x}=1-\frac{1}{2x}.$$

Substitution then results in:

$$P(Y>2X)=\int_1^3 (1-\frac{1}{2x})\frac{1}{2}dx=\left[\tfrac{1}{2}x-\tfrac{1}{4}\ln x\right]_1^3=1-\tfrac{1}{4}\ln 3 \qquad ◄$$

Because each conditional pdf is in itself a pdf, we can find the expected value in the familiar way:

Definition 5.18 (*B&E, Def. 5.4.1*)

The **conditional expected value** of Y given $X = x$ is defined as:

$$E(Y\,|\,X=x)=\begin{cases}\sum_{\text{all } y} y f_{Y|X}(y\,|\,x) & \text{when } Y \text{ discrete}\\ \int_{-\infty}^{\infty} y\, f_{Y|X}(y\,|\,x)dy & \text{when } Y \text{ continuous}\end{cases}$$

Similarly, the conditional expectation of $g(X,Y)$ given $X = x$ (where $g(\cdot,\cdot)$ is any real valued function of two variables) is defined as:

$$E(g(X,Y)\,|\,X=x)=\begin{cases}\sum_{\text{all } y} g(x,y) f_{Y|X}(y\,|\,x) & \text{when } Y \text{ discrete}\\ \int_{-\infty}^{\infty} g(x,y)\, f_{Y|X}(y\,|\,x)dy & \text{when } Y \text{ continuous}\end{cases}$$

Example 5.37 (Example 5.32 continued)

Let $f_{X,Y}(x,y)=2e^{-x-y}$ for $0 \le x \le y$. Earlier, we found the conditional pdf's:

$$f_{Y|X}(y\,|\,x)=e^{x-y} \quad \text{for } y \ge x\ (>0) \quad \text{and} \quad f_{X|Y}(x\,|\,y)=\frac{e^{-x}}{1-e^{-y}} \quad \text{for } 0 \le x \le y\ (>0).$$

The conditional expectations follow (use partial integration):

$$E(Y\,|\,X=x)=\int_x^{\infty} y\,e^{x-y}dy \overset{\text{P.I.}}{=} [-ye^{x-y}]_{y=x}^{\infty}+\int_x^{\infty} e^{x-y}dy=x+1 \quad \text{(only defined when } x>0)$$

$$E(X\,|\,Y=y)=\int_0^y x\frac{e^{-x}}{1-e^{-y}}dx=\ ..\ ..\ ..\ =\frac{e^y-y-1}{e^y-1} \quad \text{(only defined when } y>0) \qquad ◄$$

The conditional expectation $\mathrm{E}(Y \mid X = x)$ is therefore a function of x (and never of y!), and to stress that we can write this as $k(x) = \mathrm{E}(Y \mid X = x)$. Then we can see that $k(X) = \mathrm{E}(Y \mid X)$ is just a function of the random variable X, of which we can determine the expectation by using the pdf of X; so if X is continuous, we get: $\mathrm{E}[k(X)] = \int_{-\infty}^{\infty} k(x) \cdot f_X(x)\, dx$. This expectation is always equal to $\mathrm{E}(Y)$:

Theorem 5.21 (*B&E, Th.5.4.1*)

If X and Y are any two random variables, then:

$$\mathrm{E}(Y) = \mathrm{E}[\mathrm{E}(Y \mid X)]$$

<u>*Proof*</u>

(For the continuous case)

$$\begin{aligned}
\mathrm{E}[\mathrm{E}(Y \mid X)] &= \int_{-\infty}^{\infty} \mathrm{E}(Y \mid X = x) f_X(x)\, dx \\
&= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} y\, f_{Y|X}(y \mid x) dy \right] f_X(x)\, dx = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y\, f_{Y|X}(y \mid x) f_X(x)\, dx\, dy \\
&= \int_{-\infty}^{\infty} y \int_{-\infty}^{\infty} f_{X,Y}(x,y)\, dx\, dy = \int_{-\infty}^{\infty} y\, f_Y(y)\, dy = \mathrm{E}(Y).
\end{aligned}$$

This theorem can be very helpful in the determination of the (unconditional) expectation $\mathrm{E}(Y)$, when:

1 the conditional distribution of $Y \mid X = x$ is known for each x;
2 $\mathrm{E}(Y \mid X = x)$ can be determined (as a function of x);
3 the distribution of X is known as well.

<u>*Example 5.38*</u> <u>*(Example 5.34 continued)*</u>

Let $X \sim$ UNIF[1 , 3], and $Y \mid X = x \sim$ UNIF $[2x - 1 , 4x-1]$. We can now determine $\mathrm{E}(Y)$ without first having to determine the pdf of Y:

$$\mathrm{E}(Y \mid X = x) = \frac{(4x-1)+(2x-1)}{2} = 3x - 1 \quad \text{(because } Y \mid X = x \sim \text{UNIF } [2x-1 , 4x-1])$$

$$\Rightarrow \quad \mathrm{E}(Y) = \mathrm{E}(\mathrm{E}(Y \mid X)) = \mathrm{E}(3X - 1) = 3\mathrm{E}(X) - 1 = 5$$ ◄

<u>*Example 5.39*</u>

Assume that the number of spelling errors X in a thesis is POI(20) distributed. For each error, assume also that a corrector finds this with a probability of 0.85, in other words the conditional distribution of the number of errors found (Y) is $Y \mid X = x \sim$ BIN(x, 0.85), with conditional expectation $\mathrm{E}(Y \mid X = x) = 0.85x$.

The expected total number of errors found is therefore:

$$\mathrm{E}(Y) = \mathrm{E}(\ \mathrm{E}(Y \mid X) = \mathrm{E}(\ 0.85X) = 0.85\ \mathrm{E}(X) = 0.85 \cdot 20 = 17$$

Again we see that it is not necessary to first determine the distribution of Y. ◄

We can generalise the previous theorem:

Theorem 5.22 (*B&E, Th.5.4.4*)

If X and Y are two random variables and $g(\cdot,\cdot)$ is any function of two variables, then:

$$\mathrm{E}[g(X,Y)] = \mathrm{E}[\mathrm{E}(g(X,Y) \mid X)]$$

<u>*Proof*</u>

$g(X,Y)$ is a random variable itself, so this result follows directly from Theorem 5.21.

Theorem 5.23 (*B&E, Th.5.4.5*)

If X and Y are two random variables and $g(\cdot,\cdot)$ is any function of two variables, then:

$$\mathrm{E}[g(X)Y \mid X = x)] = g(x)\,\mathrm{E}[Y \mid X = x]$$

<u>*Proof*</u>

(For Y discrete): $\mathrm{E}[g(X)Y \mid X = x)] = \mathrm{E}[g(x)Y \mid X = x)] =$

$$= \sum_{\text{all } y} g(x) y f_{Y|X}(y \mid x) = g(x) \sum_{\text{all } y} y f_{Y|X}(y \mid x)$$

$$= g(x)\mathrm{E}(Y \mid X = x) = g(x)\,\mathrm{E}[Y \mid X = x].$$

<u>*Example 5.40*</u> <u>*(Example 5.31 continued)*</u>

If $(X,Y) \sim \mathrm{MULT}(n, p_1, p_2)$, then we know already that $X \sim \mathrm{BIN}(n, p_1)$, $Y \sim \mathrm{BIN}(n, p_2)$ and $f_{X,Y}(x,y) = \dfrac{n!}{x!\,y!(n-x-y)!} p_1^{\,x} p_2^{\,y} (1-p_1-p_2)^{(n-x-y)}$.

Assume we need to find the covariance, $\mathrm{Cov}(X,Y)$. Using the joint pdf, we could first determine $\mathrm{E}(XY)$, followed by the covariance. Here, we choose to do this using the conditional pdf of Y given $X = x$, because we already know that $Y \mid X = x \sim \mathrm{BIN}(n-x, \frac{p_2}{1-p_1})$ (see Example 5.31). So

$$\mathrm{E}[Y \mid X = x] = (n-x)\frac{p_2}{1-p_1}.$$

We now apply Theorem 5.22, followed by Theorem 5.23:

$$\mathrm{E}(XY) \overset{Th.5.22}{=} \mathrm{E}[\mathrm{E}(XY \mid X)] \overset{Th.5.23}{=} \mathrm{E}[X\ \mathrm{E}(Y \mid X)]$$

$$= \mathrm{E}\left[X(n-X)\frac{p_2}{1-p_1} \right] = \frac{p_2}{1-p_1}\mathrm{E}\big[X(n-X)\big] = \frac{p_2}{1-p_1}[n\mathrm{E}(X) - \mathrm{E}(X^2)]$$

Because $X \sim \mathrm{BIN}(n, p_1)$, it follows that:

$$\mathrm{E}(X) = np_1 \text{ and } \mathrm{E}(X^2) = \mathrm{Var}(X) + [\mathrm{E}(X)]^2 = np_1(1-p_1) + n^2 p_1^2 = np_1(1+(n-1)p_1).$$

Substitution yields:

$$\mathrm{E}(XY) = \frac{p_2}{1-p_1}[n^2 p_1 - np_1(1+(n-1)p_1)] = n(n-1)p_1 p_2$$

and therefore: $\mathrm{Cov}(X,Y) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y) = n(n-1)p_1 p_2 - np_1 np_2 = -np_1 p_2$ ◄

The conditional variance is simply the variance of a conditional distribution:

Definition 5.19 (*B&E, Def. 5.4.2*)

The conditional variance of Y given $X = x$ is defined as:

$$\mathrm{Var}(Y \mid X = x) = \mathrm{E}\{[Y - \mathrm{E}(Y \mid X = x)]^2 \mid X = x\} = \mathrm{E}(Y^2 \mid X = x) - [\mathrm{E}(Y \mid X = x)]^2$$

To find the variance of a random variable, we can use the concept of conditioning as well:

Theorem 5.24 (*B&E, Th.5.4.3*)

$$\mathrm{Var}(Y) = \mathrm{Var}[\mathrm{E}(Y \mid X)] + \mathrm{E}\,(\mathrm{Var}\,(Y \mid X))$$

<u>*Proof*</u>

$$
\begin{aligned}
\mathrm{E}[\mathrm{Var}(Y \mid X)] &= \mathrm{E}[\ \mathrm{E}(Y^2 \mid X) - \{\mathrm{E}(Y \mid X)\}^2\ ] \\
&= \mathrm{E}[\ \mathrm{E}(Y^2 \mid X)]\ -\ \mathrm{E}[\{\mathrm{E}(Y \mid X)\}^2\ ] \\
&= \mathrm{E}(Y^2)\ -\ \mathrm{E}[\{\mathrm{E}(Y \mid X)\}^2\ ] \\
&= \mathrm{E}(Y^2) - (\mathrm{E}(Y))^2 - \left[\mathrm{E}[\{\mathrm{E}(Y \mid X)\}^2\ ] - (\mathrm{E}(Y))^2\right] \\
&= \mathrm{Var}(Y) - \left[\mathrm{E}[\{\mathrm{E}(Y \mid X)\}^2\ ] - [\mathrm{E}\{\mathrm{E}(Y \mid X)\}]^2\right] \\
&= \mathrm{Var}(Y) - \mathrm{Var}[\mathrm{E}(Y \mid X)]
\end{aligned}
$$

Example 5.41 *(Example 5.39 continued)*

$X \sim$ POI(20), where X = the number of spelling mistakes, and $Y \mid X = x \sim$ BIN(x, 0.85) where Y = the number of spelling mistakes found by a corrector. Therefore: $\mathrm{E}(Y \mid X = x) = 0.85x$ and $\mathrm{Var}(Y \mid X = x) = 0.85 \cdot 0.15x$.

Using the Theorem above, we obtain:

$$
\begin{aligned}
\mathrm{Var}(Y) &= \mathrm{Var}(\mathrm{E}(Y \mid X)) + \mathrm{E}\,(\mathrm{Var}\,(Y \mid X)) = \mathrm{Var}(0.85X) + \mathrm{E}\,(0.85 \cdot 0.15X) = \\
&= 0.85^2\,\mathrm{Var}(X) + 0.85 \cdot 0.15\ \mathrm{E}(X)\ .
\end{aligned}
$$

Because we know the expectation and variance of a Poisson distributed random variable (so $\mathrm{E}(X) = 20$ and $\mathrm{Var}(X) = 20$, we finally get: $\mathrm{Var}(Y) = 0.85^2 \cdot 20 + 0.85 \cdot 0.15 \cdot 20 = 17$.

Remark. In this example, we might show that the marginal distribution of Y is again Poisson distributed, but now with parameter 17. However, the point here is that the determination of this distribution is not essential in order to find the expectation and the variance of Y. ◄

Multivariate conditional distributions

The concepts of conditional distributions with two random variables X and Y can be extended to k-dimensional random variables. We will just give an example here.

Example 5.42 *(Example 5.15 continued)*

Consider the joint pdf $f_{X_1,X_2,X_3}(x_1,x_2,x_3) = 6 \quad$ for $0 < x_1 < x_2 < x_3 < 1$.

We know already (see Example 5.15):

$$f_{X_1}(x_1) = 3 - 6x_1 + 3x_1^2 \qquad \text{for } 0 < x_1 < 1$$

$$f_{X_2}(x_2) = 6x_2 - 6x_2^2 \qquad \text{for } 0 < x_2 < 1$$

$$f_{X_3}(x_3) = 3x_3^2 \qquad \text{for } 0 < x_3 < 1$$

$$f_{X_1,X_2}(x_1,x_2) = 6(1 - x_2) \qquad \text{for } 0 < x_1 < x_2 < 1$$

For example, the conditional joint pdf of X_1 and X_2 given that $X_3 = x_3$ can be found by:

$$f_{X_1,X_2|X_3}(x_1,x_2 \mid x_3) = \frac{f_{X_1,X_2,X_3}(x_1,x_2,x_3)}{f_{X_3}(x_3)} = \frac{6}{3x_3^2} = \frac{2}{x_3^2} \qquad \text{for } 0 < x_1 < x_2 < x_3$$

And the conditional pdf of X_3 given that $X_1 = x_1$ and $X_2 = x_2$:

$$f_{X_3|X_1,X_2}(x_3 \mid x_1,x_2) = \frac{f_{X_1,X_2,X_3}(x_1,x_2,x_3)}{f_{X_1,X_2}(x_1,x_2)} = \frac{6}{6(1-x_2)} = \frac{1}{(1-x_2)} \qquad \text{for } x_2 < x_3 < 1$$

And the conditional pdf of X_1 given that $X_2 = x_2$ is of course:

$$f_{X_1|X_2}(x_1 \mid x_2) = \frac{f_{X_1,X_2}(x_1,x_2)}{f_{X_2}(x_2)} = \frac{6(1-x_2)}{6x_2 - 6x_2^2} = \frac{1}{x_2} \qquad \text{for } 0 < x_1 < x_2$$ ◄

Linear Regression

Like stated earlier, E(Y | $X = x$) is a function of x; this function is often called the *regression function*. A very important part of statistical theory is dedicated to estimating these functions. This starts by postulating a certain form (a formula with unknown coefficients) of the relationship between x and the expected value of Y, followed by estimating those coefficients using observations on both variables in a random sample. Very often, a linear function is assumed for E(Y | $X = x$), so E(Y | $X = x$) = a + bx; when we estimate the coefficients a and b, we are dealing with *linear regression*. (In the next section, we will see that E(Y | $X = x$) is linear whenever (X, Y) has a bivariate normal distribution).
The term regression function will also be used for the resulting equation, so the equation after estimating the coefficients. The variable X is often called the explanatory or independent variable, and Y the dependent variable. Of course, the fact that X is called independent doesn't mean that X and Y are independent random variables as defined in this chapter. But often a causal relationship is assumed, in the sense that the value of X determines to some degree the value of Y, and not the other way around. X is then seen as 'cause', and Y as 'consequence or effect'. The course Econometrics I will almost entirely be dedicated to the topis of linear regression.

5.6 Joint moment generating functions

(B&E, pages 186-188)
Moment generating functions can be defined as well for k-dimensional random variables:

Definition 5.20 (*B&E, Def. 5.5.1*)

The **joint moment generating function** for $\boldsymbol{X} = (X_1, ..., X_k)$ is defined as:

$$M_{X_1,...,X_k}(t_1,...,t_k) = \mathrm{E}\left[\exp\left(\sum_{i=1}^{k} t_i X_i\right)\right]$$

(as long as this expected value exists for values of t_i around 0, for all i)

For the bivariate situation, we have: $M_{X,Y}(t_1,t_2) = \mathrm{E}\left[e^{t_1X+t_2Y}\right]$.

In the previous course, it has been shown that the r-th moment of a random variable can be determined by differentiating the mgf r times, and substitute t by the value 0. In a similar way, we can show that

$$\mathrm{E}[X_i^r X_j^s] = \left[\frac{\partial^{r+s}}{\partial t_i^r \partial t_j^s} M_{X_1,...,X_k}(t_1,...,t_k)\right]_{t_1=0,...,t_k=0}$$

For example: $\mathrm{E}[XY] = \left[\frac{\partial^2}{\partial t_1 \partial t_2} M_{X,Y}(t_1,t_2)\right]_{t_1=0,t_2=0}$

From any given joint mgf, the marginal mgf's are very easy to obtain:

$M_X(t) = M_{X,Y}(t,0)$ en $M_Y(t) = M_{X,Y}(0,t)$

Just as was the case with univariate mgf's, the 'uniqueness property' is valid for multivariate mgf's, so if for example the pair (X_1, X_2) has the same joint mgf as (Y_1, Y_2), then (X_1, X_2) must have the same distribution as (Y_1, Y_2)).

The proof for the following theorem in one direction follows directly from Theorem 5.11 (1a) . The proof in the opposite direction is more complex and will be skipped here.

Theorem 5.25 (*B&E, Th.5.5.1*)

If $M_{X,Y}(t_1,t_2)$ exists, then the random variables X and Y are independent if and only if:

$$M_{X,Y}(t_1,t_2) = M_X(t_1) M_Y(t_2)$$

Note the large difference between Theorem 5.25 and Theorem 5.11 (1c), which was about the univariate mgf of the sum $X + Y$ of two independent random variables!!

Example 5.43

If $\boldsymbol{X} = (X_1,...,X_k) \sim \text{MULT}(n, p_1,..., p_k)$, then $f_{\boldsymbol{X}}(x_1,...,x_k) = \dfrac{n!}{x_1!x_2!...x_k!x_{k+1}!} p_1^{x_1} p_2^{x_2} ... p_{k+1}^{x_{k+1}}$.

$$M_{X_1,...,X_k}(t_1,...,t_k) = \text{E}\left[\exp\left(\sum_{i=1}^{k} t_i X_i\right)\right]$$

$$= \sum \cdots \sum \frac{n!}{x_1!x_2!...x_k!x_{k+1}!}(p_1 e^{t_1})^{x_1}(p_2 e^{t_2})^{x_2} \cdots (p_k e^{t_k})^{x_k} p_{k+1}^{x_{k+1}}$$

$$= (p_1 e^{t_1} + p_2 e^{t_2} \cdots p_k e^{t_k} + p_{k+1})^n \qquad \text{(Binomial theorem)}$$

By setting $t_3,...,t_k$ all equal to 0, we can derive the joint mgf of (X_1, X_2):

$$M_{X_1,X_2}(t_1,t_2) = M_{X_1,...,X_k}(t_1,t_2,0,...,0)$$

$$= (p_1 e^{t_1} + p_2 e^{t_2} + p_3 + \cdots + p_k + p_{k+1})^n = (p_1 e^{t_1} + p_2 e^{t_2} + (1 - p_1 - p_2))^n$$

Which in turn can be recognised as the mgf of a $\text{MULT}(n, p_1, p_2)$ distribution.

The covariance between X_1 and X_2 can be found (and certainly a lot simpler than in Example 5.40):

$$\text{E}[X_1 X_2] = \left[\frac{\partial^2}{\partial t_1 \partial t_2} M_{X_1,X_2}(t_1,t_2)\right]_{t_1=0,t_2=0} = n(n-1)p_1 p_2$$

$$\text{E}[X_1] = \left[\frac{\partial}{\partial t_1} M_{X_1,X_2}(t_1,0)\right]_{t_1=0} = np_1 \quad \text{and} \quad \text{E}[X_2] = \left[\frac{\partial}{\partial t_2} M_{X_1,X_2}(0,t_2)\right]_{t_2=0} = np_2$$

so $\text{Cov}(X_1, X_2) = n(n-1)p_1 p_2 - np_1 np_2 = -np_1 p_2$ ◄

Example 5.44

Consider the joint pdf for (X, Y) $f_{X,Y}(x, y) = 2$ voor $0 < x < y < 1$.

The joint mgf can be found by:

$$M_{X,Y}(t,u) = \text{E}(e^{tX+uY}) = \int_0^1 \int_x^1 e^{tx+uy} 2dy\, dx = \ldots.$$

(further evaluation is possible, but not the point of this example) ◄

5.7 Bivariate normal distribution

(B&E, pages 185-186, 188)

The bivariate normal distribution is the bivariate version of the normal distribution.

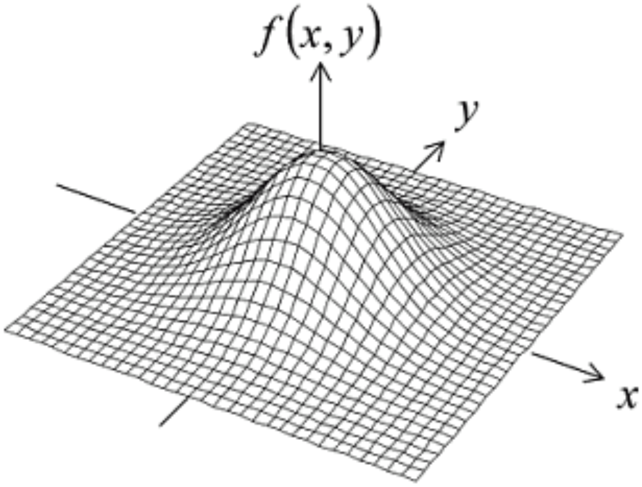
Definition 5.21 (*B&E, Eq. 5.4.11*)

The two-dimensional random variable (X,Y) has a **bivariate normal distribution** with parameters $\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2$ and ρ, notation $(X,Y) \sim \text{BVN}(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$ if the joint pdf is given by:

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right]}$$

where $\sigma_X > 0,\ \sigma_Y > 0,\ -1 \le \rho \le 1$.

Below, a 3-dimensional figure for a bivariate normal distribution is shown:



The proofs in this section are a bit more complex than most other proofs. It is sufficient to focus just the outlines of the proofs (and not on all mathematical manipulations).

Theorem 5.26 (*B&E, Th.5.4.7*)

If $(X,Y) \sim \mathrm{BVN}(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$, then $X \sim \mathrm{N}(\mu_X, \sigma_X^2)$ and $Y \sim \mathrm{N}(\mu_Y, \sigma_Y^2)$.

<u>*Proof*</u>

We start by recalling that, by definition: $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy$.

We are going to perform the substitution $v = \dfrac{y-\mu_Y}{\sigma_Y}$, so $dv = \dfrac{dy}{\sigma_Y}$ and we obtain:

$$f_X(x) = \int_{-\infty}^{\infty} \frac{\sigma_Y}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)v + v^2\right]\right\} dv$$

$$= \int_{-\infty}^{\infty} \frac{1}{2\pi\sigma_X\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - \frac{1}{2(1-\rho^2)}\left(v - \rho\frac{x-\mu_X}{\sigma_X}\right)^2\right] dv$$

As a second step, we will substitute $w = \dfrac{v-\rho(x-\mu_X)/\sigma_X}{\sqrt{1-\rho^2}}$, with $dw = \dfrac{dv}{\sqrt{1-\rho^2}}$:

$$f_X(x) = \int_{-\infty}^{\infty} \frac{\sqrt{1-\rho^2}}{2\pi\sigma_X\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - \frac{1}{2}w^2\right] dw$$

$$= \frac{1}{\sqrt{2\pi}\sigma_X} \exp\left[-\frac{1}{2}\left(\frac{x-\mu_X}{\sigma_X}\right)^2\right] \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}w^2\right] dw$$

$$= \frac{1}{\sqrt{2\pi}\sigma_X} \exp\left[-\frac{1}{2}\left(\frac{x-\mu_X}{\sigma_X}\right)^2\right]$$

This result shows that $X \sim \mathrm{N}(\mu_X, \sigma_X^2)$. And because of symmetry reasons: $Y \sim \mathrm{N}(\mu_Y, \sigma_Y^2)$.

Theorem 5.27 (*B&E, Th.5.4.8*)

If $(X,Y) \sim \mathrm{BVN}(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$, then

$$Y \mid X = x \sim \mathrm{N}(\mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X),\ \sigma_Y^2(1-\rho^2)\ ) \text{ and}$$

$$X \mid Y = y \sim \mathrm{N}(\mu_X + \rho\frac{\sigma_X}{\sigma_Y}(y-\mu_Y),\ \sigma_X^2(1-\rho^2)\ )$$

<u>Proof</u>

By definition: $f_{Y|X}(y \mid x) = \dfrac{f_{X,Y}(x,y)}{f_X(x)}$, so

$$f_{Y|X}(y \mid x) = \frac{\dfrac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}\exp\left\{-\dfrac{1}{2(1-\rho^2)}\left[\left(\dfrac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\dfrac{x-\mu_X}{\sigma_X}\right)\left(\dfrac{y-\mu_Y}{\sigma_Y}\right)+\left(\dfrac{y-\mu_Y}{\sigma_Y}\right)^2\right]\right\}}{\dfrac{1}{\sqrt{2\pi}\sigma_X}\exp\left[-\dfrac{1}{2}\left(\dfrac{x-\mu_X}{\sigma_X}\right)^2\right]}$$

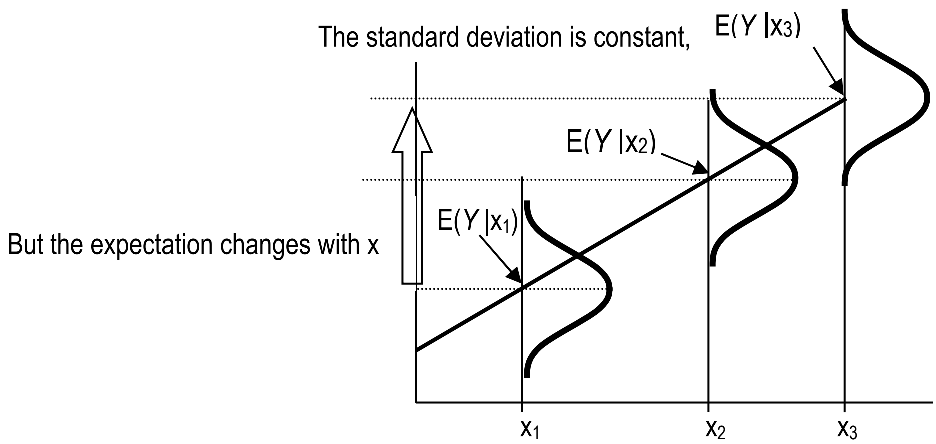
$$= \frac{1}{\sqrt{2\pi}\sigma_Y\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\rho^2\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right)+\left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right]\right\}$$

$$= \frac{1}{\sqrt{2\pi}\sigma_Y\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2\sigma_Y^2(1-\rho^2)}\left[y-\mu_Y-\rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X)\right]^2\right\}$$

This can be recognised as the pdf of a $\mathrm{N}(\mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X),\ \sigma_Y^2(1-\rho^2)\ )$ distribution.

Therefore also: $X \mid Y = y \sim \mathrm{N}(\mu_X + \rho\frac{\sigma_X}{\sigma_Y}(y-\mu_Y),\ \sigma_X^2(1-\rho^2)\ )$

This theorem implies that, when we slice the 3-dimensional figure above at any arbitrary value of *x*, the intersection will always show the well-known form of a univariate normal distribution. Furthermore, the theorem shows that the regression function $\mathrm{E}(Y \mid X = x)$ is always a linear function of *x* (with slope $\rho\frac{\sigma_Y}{\sigma_X}$), and that the conditional variance of $Y \mid X = x$ does not depend on the value of *x* (and therefore is a constant). These conclusions are actually exactly the assumptions that any simple linear regression should comply with (see also page 36).



Theorem 5.28 (*B&E, Ex.5.5.2*)

If $(X,Y) \sim \text{BVN}(\mu_X,\mu_Y,\sigma_X^2,\sigma_Y^2,\rho)$, then the joint moment generating function is:

$$M_{X,Y}(t_1,t_2) = \exp[\mu_X t_1 + \mu_Y t_2 + \tfrac{1}{2}(\sigma_X^2 t_1^2 + \sigma_Y^2 t_2^2 + 2\rho\sigma_X\sigma_Y t_1 t_2)]$$

Proof

$$\begin{aligned} M_{X,Y}(t_1,t_2) &= \text{E}\left[e^{t_1X+t_2Y}\right] \\ &= \text{E}\left\{\text{E}\left[e^{t_1X+t_2Y} \mid X\right]\right\} && \text{(because of Theorem 5.22)} \\ &= \text{E}\left\{e^{t_1X}\text{E}\left[e^{t_2Y} \mid X\right]\right\} && \text{(Theorem 5.23 applied to part within curly brackets)} \end{aligned}$$

Using Theorem 5.27 and the general formula for the mgf of a univariate normal distributed variable, it follows that

$$\text{E}\left[e^{t_2Y} \mid X = x\right] = M_{Y|X=x}(t_2) = \exp\left\{\left[\mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X)\right]t_2 + \sigma_Y^2(1-\rho^2)t_2^2/2\right\}$$

When we substitute this result into $\text{E}\left\{e^{t_1X}\text{E}\left[e^{t_2Y} \mid X\right]\right\}$, we obtain with some effort the desired result.

The bivariate normal distribution has five parameters. The meaning of μ_X,μ_Y,σ_X^2 and σ_Y^2 are as usual. As you might expect, the fifth parameter, ρ, represents the coefficient of correlation of *X* and *Y*. We will check that now using the joint mgf. First, we will determine E(*XY*):

$$\begin{aligned} \text{E}[XY] &= \left[\frac{\partial^2}{\partial t_1 \partial t_2}M_{X,Y}(t_1,t_2)\right]_{t_1=0,t_2=0} \\ &= \big[\exp[\mu_X t_1 + \mu_Y t_2 + \tfrac{1}{2}(\sigma_X^2 t_1^2 + \sigma_Y^2 t_2^2 + 2\rho\sigma_X\sigma_Y t_1 t_2)] \\ &\qquad \times[(\mu_Y + \sigma_Y^2 t_2 + \rho\sigma_X\sigma_Y t_1)(\mu_X + \sigma_X^2 t_1 + \rho\sigma_X\sigma_Y t_2) + \rho\sigma_X\sigma_Y]\big]_{t_1=0,t_2=0} \\ &= \mu_X\mu_Y + \rho\sigma_X\sigma_Y \end{aligned}$$

$$\Rightarrow \quad \text{Cov}(X,Y) = \text{E}[XY] - \text{E}(X)\text{E}(Y) = \mu_X\mu_Y + \rho\sigma_X\sigma_Y - \mu_X\mu_Y = \rho\sigma_X\sigma_Y$$

$$\Rightarrow \quad \rho_{X,Y} = \frac{\text{Cov}(X,Y)}{\sigma_X\sigma_Y} = \rho$$

If *X* and *Y* are independent, then the parameter ρ is of course equal to 0. Vice versa, if ρ = 0, then the joint pdf of the bivariate normal distribution can be written as the product of the two marginal pdf's. Using Theorem 5.10, we can therefore conclude that *X* and *Y* are independent. In other words: for bivariate normal random variables, 'uncorrelated' has exactly the same meaning as 'independent', which is, as we have seen earlier, not true in general.

All the above can be extended to *k*-dimensional multivariate normal distributions.

Definition 5.22 (*B&E, Def. 15.4.1*)

The *k*-dimensional random variable $\boldsymbol{X} = (X_1,\ldots,X_k)$ has the **multivariate normal distribution** if the joint pdf is given by:

$$f_{\boldsymbol{X}}(x_1,\ldots,x_k) = \frac{1}{(2\pi)^{k/2}\,|\Sigma|^{1/2}}e^{-\frac{1}{2}(\boldsymbol{x}-\mu)\Sigma^{-1}(\boldsymbol{x}-\mu)^T}$$

where Σ is the variance-covariance matrix,as discussed on page 27 and $\mu = (\mu_{X_1},\ldots,\mu_{X_k})$.

5.8 Problems

5.1 Given the pdf $f_{X,Y}(x, y) = (x + 2y) / c$ for discrete x, y with $x = 0, 1, 2$, and $y = 0, 1, 2$ (0 otherwise). Determine the marginal pdf of X and Y.

5.2 Throw two unbiased tetrahedra, that is, 'dices' with four sides (faces). Let the numbers respectively be X and Y. Construct the joint probability table for X and $U = |X - Y|$. Determine also the marginal pdf of X and U, and derive E(X) and E $|X - Y|$.

5.3 (B&E 4.8) The discrete random variables X and Y have the joint pdf:

$$f_{X,Y}(x, y) = c\frac{2^{x+y}}{x!\,y!} \qquad x = 0,1,2,.... \; y = 0,1,2,...$$

a. Determine the constant c.
b. Determine the marginal pdf of X and Y.

5.4 Let the joint pdf of (X, Y) be given by

$P(X = x\,,\, Y = y) = k \qquad$ for $x = 1, 2, \ldots, 10$ and $y = (10 - x), \ldots, 9, 10$

a Calculate k.
b Determine the marginal pdf of X and Y.

5.5 In a large population 60% favour a certain measure and 30% disapprove. The remaining 10% has no opinion in the matter. So, these are three disjoint categories. Ten persons are randomly chosen from this population. Determine the probability that:

a the sample consists of six persons in favour, three persons against and one without an opinion.
b the sample consists of four persons in favour, four persons against and two without an opinion.

5.6 (B&E 4.6) An unbiased die is thrown 12 times. If X_1 denotes the number of ones, X_2 the number of twos, etc., determine the probability that

a $(X_1 = 2, X_2 = 3, X_3 = 1, X_4 = 0, X_5 = 4, X_6 = 2)$
b $(X_1 = X_2 = X_3 = X_4 = X_5 = X_6)$
c $(X_1 = 1, X_2 = 2, X_3 = 3, X_4 = 4)$
d Give an expression for the joint pdf of X_1, X_3 and X_5

5.7 A card player draws randomly (without replacement) 13 cards from a complete deck with 52 cards. Suppose the player is interested in the joint distribution of the number of the four different suits (so the number of spades, number of hearts, number of diamonds and number of clubs). Define a suitable multi-dimensional random variable that describes this the best and give the associated probability distribution function. Calculate the probability that the player draws 2 spades, 4 clubs, 1 heart and 6 diamonds.

5.8 A student takes a multiple-choice test consisting of n_1 basic questions and n_2 more difficult questions. Let X denote the number of correctly answered basic questions and Y the number correctly answered difficult questions.

a Suppose that the probability of correctly answering any of these questions equals p.
  a1 Determine the joint probability distribution of X and Y.
  a2 Determine the probability distribution of $X + Y$.
b Suppose that the probability of correctly answering a basic questions equals p, but for difficult questions the probability is different, say p_1.
  b1 Determine again the joint probability distribution of X and Y.
  b2 Determine the marginal distributions of X and Y.
  b3 Is $X + Y$ binomial again, and if so, what is the 'probability of success'?

5.9 The simultaneous pdf of X and Y is

$f_{X,Y}(x, y) = k\,x\,y \quad$ for $\quad 0 \le x < 2\,, \quad 0 \le y < 1\,, \quad 2y < x$

a Draw the support of the distribution.
b Determine the constant k.
c Calculate $P(Y < X/3)$.

5.10 (B&E 4.20a) The simultaneous density of X and Y is given by: $f_{X,Y}(x, y) = 8xy \quad$ for $0 < x < y < 1$.

Determine the simultaneous CDF of X and Y.

5.11 (B&E 4.18) For a pair of random variables X and Y the following CDF is given:

$$F_{X,Y}(x,y)=\begin{cases}0.5xy(x+y) & 0<x<1,\ 0<y<1\\ 0.5x(x+1) & 0<x<1,\ 1\le y\\ 0.5y(y+1) & 1\le x,\ 0<y<1\\ 1 & 1\le x,\ 1\le y\end{cases}$$

and 0 otherwise. Determine:

a the simultaneous pdf

b $P(X\le 0.5, Y\le 0.5)$

c $P(X<Y)$

5.12 (B&E 4.12) Consider the function $F_{X,Y}(x, y)$ defined as:

$$F_{X,Y}(x,y)=\begin{cases}0.25(x+y)^2 & \text{for } 0\le x<1, 0\le y<1\\ 0 & \text{for } x<0 \text{ of } y<0\\ 1 & \text{otherwise}\end{cases}$$

Is this a bivariate CDF?

5.13 The simultaneous density of X and Y is

$f_{X,Y}(x, y) = k$ for $0\le x<5$, $0\le y<9$, $y < x^2$

a Determine the value of k. Does the function above indeed represent a probability density function?

b Determine the probability that $Y > X$.

5.14 Determine the marginal density of X based on the simultaneous density from the previous exercise.

5.15 Let a continuous simultaneous density be given by

$f_{X,Y}(x,y)=k$ for $1 < x < 2$, $2 < y < 4$

a Determine k.

b Determine the marginal densities of the corresponding variables X and Y.

5.16 A continuous simultaneous pdf is given by $f_{X,Y}(x, y) = 1$ at the support set defined by $x, y \ge 0$ and $2x + y \le 2$ (0 otherwise).
Determine the marginal densities of X and Y.

5.17 The random variables X and Y have a simultaneous distribution defined by

$f_{X,Y}(x, y) = cy$ at the support set: $0<x<4$ and $0<y<1-\frac{1}{4}x$

a Determine c.

b Determine the marginal distribution of X and use it to find the expected value of X.

c Determine the marginal distribution of Y and use it to find the expected value of Y.

5.18 Three random variables X, Y and Z have the following simultaneous pdf:

$f_{X,Y,Z}(x,y,z)=c(x+y)e^{-x-y-z}$ for $x,y,z>0$

Show that $c = ½$ and determine the marginal pdf of Y.

5.19 Use the table constructed at problem 5.2. Determine the probability function of $W = X + |X - Y|$. Show that $E(X + |X - Y|)$, which can be derived from this probability distribution, equals $E(X) + E|X - Y|$.

5.20 Throw with an unbiased tetrahedron and denote the number of dots X. Next, throw X (i.e. 1, 2, 3 or 4) times with a fair coin. Denote the number of times ‘head’ as Y.

a Construct the simultaneous probability table for (X, Y).

b Show that $E(X + Y) = E(X) + E(Y)$.

c Show that $\mathrm{Var}(X + Y) \ne \mathrm{Var}(X) + \mathrm{Var}(Y)$.

5.21 The simultaneous density of X and Y is given by: $f_{X,Y}(x,y)=8xy$ for $0<x<y<1$.

Determine $E(X^iY^j)$ where i and j denote non-negative integers.

Also determine $E(X+Y)$ using the simultaneous density and verify that $E(X + Y) = E(X) + E(Y)$.

5.22 Investigate for all previous exercises, in which we are dealing with an X and a Y variable, whether the mentioned random variables X and Y are independent.

5.23 In the table given below with the marginal distributions of X and Y, denoting the sales of item R and S respectively, the expected sales are 2 and 1 respectively.

		Number of items R sold				
		0	1	2	3	
Number of items S sold	0					0.2
	1					0.6
	2					0.2
	Total	0.09	0.22	0.29	0.4	1

Determine the simultaneous probabilities given that X and Y are independent.

5.24 Assume that the independent random variables X and Y are uniform distributed on [0, 5] and [1, 3] respectively. Determine the joint pdf and the joint CDF.

5.25 Suppose the independent random variables X and Y are exponential distributed with expected values 2 and 3 respectively. Determine the simultaneous density and the simultaneous distribution function.

5.26 Consider the independent random variables X and Y with simultaneous density $3e^{-6y}$ on the support $[0 , 2] \times (0 , \infty)$. Determine <u>without integration</u> the marginal pdf's of X and Y.

5.27 For this exercise, see Theorem 5.9. Derive for the continuous case that from the equality $F_{X,Y}(x,y) = F_X(x) \cdot F_Y(y)$ for all x, y, follows that $f_{X,Y}(x,y) = f_X(x) \cdot f_Y(y)$ for all x, y.

5.28 (B&E 5.1) Let X_1, X_2, X_3 and X_4 be independent random variables, each with the same distribution with expected value 5 and standard deviation 3. Define $Y = X_1 + 2X_2 + X_3 - X_4$. Determine E($Y$) and Var($Y$).

5.29 Prove Theorem 5.13.

5.30 In the table given below showing the simultaneous distribution of X and Y, denoting the sales of items H and G respectively, the expected sales are 2 and 1 respectively.

		Number of items H sold				
		0	1	2	3	Total
Number of items G sold	0	0.06	0.04	0.04	0.06	0.2
	1	0.03	0.17	0.23	0.17	0.6
	2	0	0.01	0.02	0.17	0.2
	Total	0.09	0.22	0.29	0.4	1

Calculate the covariance by using the formula $E(X - EX)(Y - EY)$. Verify the equality of the formulas in Definition 5.14 and Theorem 5.12.

5.31 Show that for arbitrary random variables X, Y, Z and W:
a $\text{Cov}(X + Y, Z + W) = \text{Cov}(X, Z) + \text{Cov}(X, W) + \text{Cov}(Y, Z) + \text{Cov}(Y, W)$
b $\text{Cov}(X + Y, X - Y) = \text{Var}(X) - \text{Var}(Y)$

5.32 One throws an unbiased die twice and counts X and Y dots, respectively.
a Calculate the covariance of X and $X + Y$.
b Calculate the covariance of $X - Y$ and $X + Y$.
c Are $X - Y$ and $X + Y$ independent?
d Is $\text{Var}(X + Y)$ equal to $\text{Var}(2X)$?

5.33 One throws an unbiased die twice and counts the number of ones (X) and sixes (Y). Note that the outcomes 2 up to 5 can be considered as one category. Calculate the covariance of X and Y
a by constructing the table with simultaneous probabilities;
b by using Theorem 5.14. [Note that $X + Y$ also has a binomial distribution, of which the variance is known!]

5.34 From a finite population {1, 2, 3, 4, 5} one draws successively two numbers X and Y without replacement.
a Find the joint pdf and calculate the covariance.
b Find the mean and variance of $X + Y$.
c Find the probabilities $P(X \le 2)$, $P(X + Y \le 5)$ and conditional probability $P(X \le 2 \mid X + Y \le 5)$.

5.35 (B&E 5.4) The joint pdf of the discrete random variables *X* and *Y* is given by:

$f_{X,Y}(x,y) = 4/(5xy)$ voor $x = 1,2$ en $y = 2,3$.

Find the covariance of *X* and *Y*.

5.36 (B&E 5.7) Suppose *X* and *Y* are independent random variables with E(X) = 2, E(Y) = 3, Var(X) = 4 and Var(Y) = 16.

a Find E($5X - Y$)
b Find Var($5X - Y$)
c Find Cov($3X + Y$, Y)
d Find Cov(X, $5X - Y$)

5.37 (B&E 5.8) Show that $\mathrm{Cov}(\sum_{i=1}^{k} a_i X_i, \sum_{j=1}^{m} b_j Y_j) = \sum_{i=1}^{k}\sum_{j=1}^{m} a_i b_j \mathrm{Cov}(X_i, Y_j)$

5.38 Use the result of the previous exercise to prove Theorem 5.18.

5.39 Calculate the correlation coefficient of the discrete random variables *X* and *Y* when the probabilities are given by

	y=0	y=1
x=0	$1/6$	$1/3$
x=1	$1/3$	$1/6$

5.40 Express ρ(aX, $bY + c$) in terms of ρ(X, Y), a, b and c.

5.41 Var(X) = 10, Var(Y) = 20 and $\rho(X, Y) = ½$. Calculate Var($2X - 3Y$).

5.42 One throws an unbiased die *n* times. The number of ones in these trials is denoted by *X*, and the number of sixes by *Y*. So, the variables *X* and *Y* have a simultaneous distribution.

a Using the binomial formulas, find E(X), E(Y), Var(X), Var(Y), E($X + Y$) and Var($X + Y$). [Note that: $X + Y \sim$ Bin(n, ⅓)]
b Use the results of a to determine the covariance of *X* and *Y*.
c Find Var($X - Y$) using the covariance found at b.
d Calculate ρ(X, Y).

5.43 (B&E 5.3) The simultaneous density of *X* and *Y* is given by:

$f_{X,Y}(x,y) = 24xy$ for $0 < x, 0 < y, x + y < 1$.

a Find the covariance and the correlation coefficient of *X* and *Y*.
b Find Cov($X + 1$, $5Y - 2$)

5.44 (B&E 5.11) The joint density of *X* and *Y* is given by:

$f_{X,Y}(x,y) = 6x$ for $0 < x < y < 1$.

Find the covariance and the correlation coefficient of *X* and *Y*.

5.45 Calculate ρ when the simultaneous pdf of *X* and *Y* (both continuous) is:

$f(x, y) = x + y$ for $0 < x < 1$, $0 < y < 1$

5.46 a Prove that for two independent variables *X* and *Y* both the covariance and correlation coefficient are 0.
b Show that the reverse does not hold always by working out the following two counter-examples:
b1 the variables *X* and *Y* have a continuous uniform distribution on the square with: $|x| + |y| \le 1$
b2 the variables *X* and *Y* have a discrete uniform distribution on four grid points: (1 , 0) (0 , 1) (−1 , 0) (0 , −1)

5.47 Let X_1, X_2 and X_3 be three uncorrelated random variables, each with the same variance σ^2. Find the correlation coefficient between $X_1 + X_2$ and $X_2 + X_3$.

5.48 Toss a fair coin twice and define $X = 0$ if the first coin turns up 'tails' and $X = 1$ for 'heads'. Define the variable *Y* analogously for the second coin. A third random variable *Z* is equal to 1 if precisely one coin turns up 'heads', and 0 otherwise. Determine the joint pdf of (*X*, *Y*, *Z*) (enumerating all possible outcomes with corresponding probabilities suffices). Are these random variables pairwise independent? Is it possible to determine the value of an arbitrary random variable given the values of the other two variables? Are (*X*, *Y*, *Z*) mutually independent?

5.49 Consider a discrete probability distribution with $f_{X,Y}(x, y) = (x + 2y) / 27$ for integer values x, y with x = 0, 1, 2, and y = 0, 1, 2. Find the conditional pdf's of X and Y (so conditional on the other r.v.) Check that $f_{X|Y}(x\,|\,2)$ is a proper pdf.

5.50 Draw 8 cards from a standard deck containing 52 cards. Let X be the number of 'jacks' and Y the number of 'aces'.
- a Give the joint pdf of X and Y.
- b Determine the conditional pdf of X given $Y = y$.
- c What is the probability of getting 3 jacks, given that 2 of the 8 cards are aces?

5.51 Suppose one spins a well-balanced Wheel of Fortune three times, and every time one marks the exact location where the wheel has stopped. Determine the probability that all three marks are within one half of the circle (so within 180 degrees). Tip: Use the location of the first mark as reference point. Let X be the location of the second mark (in radians) and let A be the event that all three marks are on the same half of the circle. Next, determine the probability of A by conditioning on the location of X, using

$$P(A) = \int_{-\infty}^{\infty} P(A\,|\,X = x) f_X(x)\,dx \qquad \text{(See Example 5.35 above).}$$

5.52 (B&E 4.20 b-e) Suppose the joint pdf of X and Y is given by:

$$f_{X,Y}(x, y) = 8xy \quad \text{for } 0 < x < y < 1 .$$

- a Find the conditional probability densities.
- b Find P ($X < 0.5 \mid Y = 0.75$).
- c Find P ($X < 0.5 \mid Y < 0.75$).

5.53 (B&E 4.30) Suppose the simultaneous density of X and Y is given by:

$$f_{X,Y}(x, y) = 60x^2 y \quad \text{for } 0 < x, 0 < y, x + y < 1 .$$

- a Find the conditional probability density of Y given $X = x$.
- b Find P ($Y > 0.1 \mid X = 0.5$).

5.54 The simultaneous density of X and Y is

$$f_{X,Y}(x, y) = \begin{cases} c(x + y) & 0 < x < 1, \quad 0 < y < 1 \\ 0 & \text{otherwise} \end{cases}$$

- a Show that $c = 1$.
- b Determine the marginal and conditional densities. Verify that for these densities the integral over their respective supports indeed equals 1.

5.55 The independent random variables X and Y have a BIN(2, ½)-and a BIN(4, ½)-distribution, respectively. Show that X given $X + Y = 3$ has a HYPERG(3, 2, 6)-distribution.

5.56 The three random variables X, Y and Z have the following simultaneous pdf:

$$f_{X,Y,Z}(x, y, z) = x^2 e^{-x(1+y+z)} \qquad \text{for } x, y, z > 0$$

- a Find the marginal pdf of Y.
- b Calculate the conditional pdf of Y and Z, given X=x, i.e. $f_{Y,Z|X}(y, z\,|\,x)$.

5.57 One throws two (distinguishable) dice. The number of dots for the first is denoted by X, the number of dots for the second by Y, the total number by Z (= sum $X + Y$).
- a Find the conditional distribution of Z given $X = 1$.
- b Find the conditional distribution of Z given $X = 6$.
- c Find the conditional distribution of X given $Z = 4$.
- d Find the conditional distribution of X given $Z = 8$.

5.58 (B&E 5.16) The simultaneous density of X and Y is given by:

$$f_{X,Y}(x, y) = e^{-y} \quad \text{for } 0 < x < y < \infty .$$

Find E($X \mid Y = y$).

5.59 Determine the joint pdf of (X, Y) when X has an EXP(1)-distribution, and the conditional distribution of Y given $X = x$ is a uniform distribution on (0, x).

5.60 The joint pdf of X and Y is
$f_{X,Y}(x, y) = (x + y) / 21$ for $x = 1, 2, 3;\ y = 1, 2$.

a Show that $f_{X,Y}(x, y)$ is indeed a proper pdf.
b Find the conditional pdf's of X given $Y = y$ and of Y given $X = x$.
c Determine the (conditional) expectation of the conditional probability distributions above.
d Find $P(Y - X \le 0)$.

5.61 Find the joint pdf of (X, Y) when Y has a BIN(3, ½)-distribution, and the conditional distribution of X given $Y = y$ is a Poisson-distribution with mean y.
Find also the expected value of X.

5.62 The simultaneous density function of X and Y is
$f_{X,Y}(x, y) = k\,x\,y$ for $0 \le x,\ 0 \le y,\ x^2 + y^2 \le 1$

a Calculate k.
b Find $E(Y \mid X = x)$ and $\mathrm{Var}(Y \mid X = x)$
c Calculate $E(Y)$ using the results of b.
d Calculate $P(Y > 0.4 \mid X = 0.5)$

5.63 (B&E 5.17) Suppose that $X \sim$ EXP(1), and that the conditional distribution of Y given $X = x$ is Poisson with mean x.
a Find $E(Y)$.
b Find $\mathrm{Var}(Y)$.

5.64 Given is the following joint pdf of the random variables X and Y

$$f_{X,Y}(x,y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y-x)^2 - x} \quad \text{for } 0 < x < \infty,\ -\infty < y < \infty$$

a Show that X has an exponential distribution (explain each step).
b Find the conditional pdf of Y given $X = x$. What kind of distribution is it?
c There exists no simple expression for the marginal pdf of Y. Instead, use $E(Y) = E(E(Y \mid X))$ to determine the expected value of Y.

5.65 Given is the following joint pdf of the random variables X, Y and Z.

$$f_{X,Y,Z}(x,y,z) = \begin{cases} ce^{-z} & \text{for } 0 < x < y < z \\ 0 & \text{otherwise} \end{cases}$$

a Find c and the marginal pdf of Y. Which well-known distribution has Y?
b Find the conditional pdf of X given $Y = y$.
c Show that $E(X \mid Y = y) = y/2$.
d *Use the results of (a) and (c)* to determine successively $E(X)$, $E(XY)$ and $\mathrm{Cov}(X,Y)$.

5.66 (B&E 5.18) One box contains five red and six black balls. A second box contains ten red and five black balls. One ball is drawn from the first box and placed in the second box. Next, two balls are drawn from the second box without replacement. What is the expected number of red balls drawn from the second box?

5.67 (B&E 5.19) The number of times a batter gets to bat in a game (denoted by N) follows a binomial distribution: $N \sim \mathrm{Bin}(6, 0.8)$. Let X denote the number of times the batter hits the ball. For each bat the batter has a probability 0.3 of hitting the ball.
a Find $E(X)$.
b Find $\mathrm{Var}(X)$.

5.68 (B&E 5.20) Let X be the number of customers arriving in a given minute at a certain ATM machine, and let Y be the number who make withdrawals. Assume that X is Poisson distributed with expected value $E(X) = 3$, and that the conditional expectation and variance of Y given $X = x$ are: $E(Y \mid X = x) = x/2$ and $\mathrm{Var}(Y \mid X = x) = (x+1)/3$.
a Find $E(Y)$.
b Find $\mathrm{Var}(Y)$.
c Find $E(XY)$.

5.69 The joint pdf of X and Y is given by: $f_{X,Y}(x, y) = k\,x\,y$ for $0 \le x < 1,\ x^2 \le y < x$
a Find the conditional expectations.
b Calculate $P(X > ½ \mid Y = ⅓)$.

5.70 (Exercise 5.49 continued). For a discrete pdf $f_{X,Y}(x, y) = (x + 2y)/27$ for $x = 0, 1, 2$, and $y = 0, 1, 2$, Find the conditional expectation Y given $X = x$.

5.71 Suppose the independent random variables X and Y are uniform distributed on [0, 5] and [1 , 3], respectively. Find the simultaneous moment generating function of (X, Y) and the mgf of $X + Y$.

5.72 (B&E 5.21) The simultaneous density of X and Y is given by:

$$f_{X,Y}(x, y) = 2e^{-x-y} \quad \text{voor } 0 < x < y .$$

Find the simultaneous moment generating function of X and Y.

5.73 (B&E 5.22) The joint pdf of X and Y is given by:

$$f_{X,Y}(x, y) = e^{-y} \quad \text{for } 0 < x < y .$$

Find the joint moment generating function of X and Y.
Next, use this mgf to determine the marginal distributions of X and of Y.

5.74 Suppose the independent random variables X and Y are exponential distributed with mean 2 and 3, respectively. Find the moment generating functions of (X, Y) and of $X + Y$.

5.75 Let X_1, X_2, and X_3 be independent random variables, with $X_i \sim \text{EXP}(\lambda)$. Let $V = X_1 - X_2$ and $W = X_3 - X_2$. Find the simultaneous moment generating function of V and W. Are V and W independent?

5.76 Given are two independent bivariate random variables: (X_1, Y_1) and (X_2, Y_2), both BVN(0, 0, 1, 1, ½)-distributed. Let $V = aX_1 + X_2$ and $W = -aY_1 + Y_2$.

- a Find the simultaneous moment generating function of V and W. For which value of a are V and W independent?
- b Find E($W \mid V = 5$) and Var($W \mid V = 100$).

5.77 Write out the joint pdf of (X , Y) when $X \sim N(0 , 1)$ and $Y \mid X = x \sim N(10 + 2x , \sigma^2)$. What distribution do we get?

5.78 Prove Theorem 5.26 using the moment generating function of X and Y.

5.79 Show that Definition 5.22 for $k = 1$ leads to the well-known pdf of a normal distribution, and for $k = 2$ to the pdf of a bivariate normal distribution.

6 Functions of multivariate random variables

This chapter will focus on the distributions of functions of multivariate random variables. In the first section, we will discuss the three different techniques which can help us finding those distributions. This will be followed by a discussion on the distribution of sums of random variables, and on the distributions and properties of a number of important sample statistics. The last section is dedicated to the special problem of finding distribution in ordered samples.

6.1 Three methods for transformations

In chapter 4 in the previous course, we already discussed methods for finding the distributions of functions of a single, univariate random variable. Three methods have been discussed there: the CDF-method, the transformation method and the moment generating function-method. These methods will now be generalised to functions of multivariate random variables.

6.1.1 The CDF-method

(B&E, pages 194-197)

In section 4.6.1 we have seen how a function $h(X)$ of a random variable X is in itself a new random variable, $W = h(X)$, of which we could write the CDF as follows:

$$F_W(w) = P(W \le w) = P(h(X) \le w)$$

Often, we will be able to express the CDF of W in terms of the CDF of X. And if X is continuous, then the pdf can then be found by determining the derivative of the CDF.

<u>*Example 6.1*</u>

Let $X \sim \text{EXP}(1)$. Consider the random variable $W = \theta\sqrt{X}$ with $\theta > 0$.

Thus: $F_W(w) = P(W \le w) = P(\theta\sqrt{X} \le w) = P(X \le (w/\theta)^2) = F_X((w/\theta)^2) = 1 - e^{-(w/\theta)^2}$ $\quad (w \ge 0)$

This gives the pdf: $f_W(w) = \dfrac{d(1 - e^{-(w/\theta)^2})}{dw} = \dfrac{2we^{-(w/\theta)^2}}{\theta^2}$ $\quad (w \ge 0,\ 0 \text{ elsewhere})$

This is the pdf of a Weibull-distribution (see section 4.5.6) with parameters θ and β = 2. ◄

The same idea still holds for a function of two random variables. The CDF of $W = h(X,Y)$ can be determined by applying the definition of a CDF:

$$P(W \le w) = P(h(X,Y) \le w)$$

The most difficult part of this method is usually the determination of the limits of the double integral, which sometimes even needs to be split into two or more parts. The exact way to do so is different from case to case. If $h(X,Y) = X + Y$, then this is all relatively easy:

$$P(X + Y \le w) = P(X \le w - Y) = \int_{-\infty}^{\infty} \left(\int_{-\infty}^{w-y} f_{X,Y}(x,y)\, dx \right) dy$$

But as soon as in the integral above we substitute $f_{X,Y}(x,y)$ by a specific formula of the pdf, we must make sure that the new integration limits reflect properly the support set of X and Y. In principle, the order of integration is arbitrary; however, it is possible that one specific order has computational advantages over the other order.

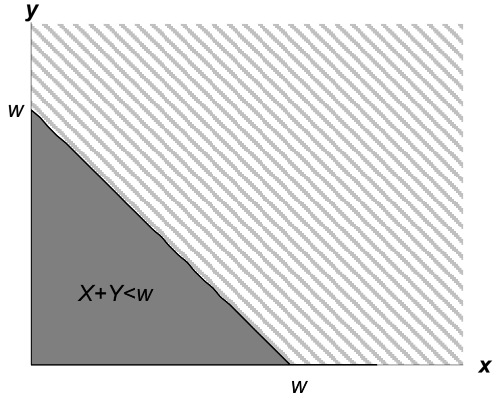
Drawing a figure can be a great help for the determination of the proper integration limits. In this figure, we will clearly indicate the support set, but also the area for which the condition $h(x, y) \le w$ holds (for some value of w). In the figures below, the support set is indicated by the diagonally shaded area, and the condition $h(x, y) \le w$ for an arbitrary value of w by the vertically shaded area. The intersection of both areas has been made visible by the solid grey area (and thus reflects the area over which the integration should be performed).

Example 6.2

Suppose X, Y are two independent random variables, each with an exponential distribution with intensity λ, so with an expected value of λ^{-1}. Thus we get:

$$f_{X,Y}(x,y) = (\lambda e^{-\lambda x})(\lambda e^{-\lambda y}) = \lambda^2 e^{-\lambda(x+y)} \quad \text{for } x, y > 0.$$

We will determine now the CDF of the sum $W = X + Y$.



Note that although the solid grey area in the figure above becomes larger for increasing value of w, the form of the area will always remain the same. We can see from the shape of this area that the variable x can range from 0 to w, and that subsequently (so for a given value of x) the variable y can range from 0 to $w - x$ (thus ensuring that always $x + y \le w$):

$$F_W(w) = P(X + Y \le w) = \int_0^w \Big(\int_0^{w-x} \lambda^2 e^{-\lambda(x+y)}\, dy \Big)\, dx = \int_0^w \Big[-\lambda e^{-\lambda(x+y)} \Big]_{y=0}^{y=w-x} dx =$$

$$= \int_0^w \Big(-\lambda e^{-\lambda w} + \lambda e^{-\lambda x} \Big)\, dx = \Big[-\lambda e^{-\lambda w} x - e^{-\lambda x} \Big]_0^w = 1 - e^{-\lambda w} - \lambda w e^{-\lambda w} \qquad (\text{for } w \ge 0)$$

By differentiating, we find the pdf :

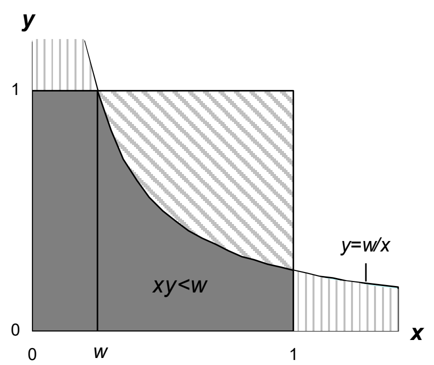
$$f_W(w) = 0 + \lambda e^{-\lambda w} - \lambda(e^{-\lambda w} - \lambda w e^{-\lambda w}) = \lambda^2 w e^{-\lambda w} \quad \text{for } w \ge 0.$$

We can recognise this result as the pdf of the Gamma-distribution with parameters 1/λ and 2.

Remark. This result could have been found in a much simpler way by using moment generating functions (see also Theorem 5.11 (1c)). ◄

Example 6.3

Consider two independent random variables X and Y , both uniformly and independently distributed on the interval (0, 1). So $f_{X,Y}(x,y) = 1 \quad \text{for } 0 \le x, y \le 1$ (check). We will find the distribution of $W = X \cdot Y$. Firstly, it is easy to see that W can assume values from 0 to 1. Now, for any arbitrary value of w between 0 and 1, the integration area is shown in the figure below. We can see that for any value of x between 0 and w, the variable y can range from 0 to 1, while for values of x between w and 1, y can range only from 0 to w/x.



Therefore, we will split the integral into two parts:

$$F_W(w) = P(X\,Y \le w) =$$

$$\int_0^w \int_0^1 1\, dy\, dx + \int_w^1 \int_0^{w/x} 1\, dy\, dx$$

$$= \int_0^w 1\, dx + \int_w^1 w/x\, dx = w - w \ln w \quad (0 \le w \le 1)$$

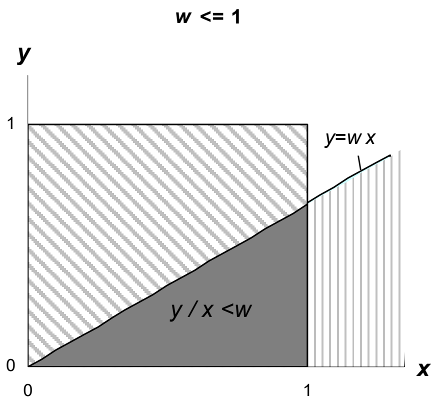
Thus:

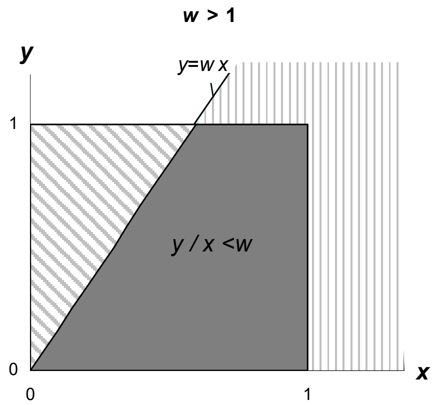
$$f_W(w) = -\ln w \quad (\text{for } 0 \le w \le 1)$$ ◄

It is also possible that the integration area changes form depending on the value of w:

Example 6.4

Again, consider two independent random variables X and Y, both uniformly and independently distributed on the interval (0, 1). Now, we will try to find the distribution of $W = Y / X$. Note that W can now assume any value from 0 to infinity. But we must now make a distinction between values of w between 0 and 1, and values larger than 1 (because the solid areas below have a different form)





$$F_W(w) = P(Y / X \le w) = \begin{cases} \displaystyle\int_0^1 \int_0^{wx} 1\, dy\, dx = \int_0^1 wx\, dx = \tfrac{1}{2} w & \text{(for } w \le 1) \\ \displaystyle\int_0^1 \int_{y/w}^1 1\, dx\, dy = \int_0^1 (1 - y/w)\, dy = 1 - \frac{1}{2w} & \text{(for } w > 1) \end{cases}$$

So the pdf becomes: $f_W(w) = \begin{cases} \tfrac{1}{2} & \text{(for } w \le 1) \\ \tfrac{1}{2} w^{-2} & \text{(for } w > 1) \end{cases}$ ◄

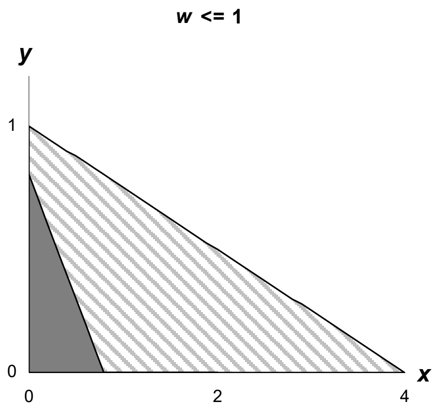
Of course, it can always be even more complicated:

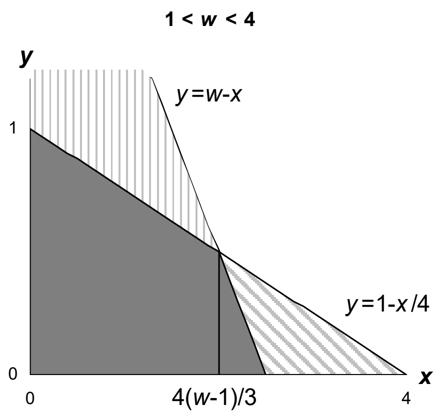
Example 6.5

Assume the random variables X and Y have the joint pdf

$$f_{X,Y}(x, y) = \frac{3y}{2} \qquad \text{on the area: } 0 < x < 4 \text{ and } 0 < y < 1 - \tfrac{1}{4} x$$

Determine now the pdf of $W = X + Y$. When drawing a figure, it becomes clear that a distinction should be made between either $0 \le w < 1$ and $1 < w < 4$ (see below). In the latter case, the integral will have to be split into two different parts. Before we can continue, we first need the intersection of the lines $y = 1 - \frac{1}{4} x$ and $w = x + y$. This gives the x-coordinate $4(w - 1) / 3$.





Case I ($0 < w \le 1$): $P(X+Y \le w) = \int_0^w \int_0^{w-x} \frac{3y}{2} dy dx = \int_0^w \frac{3}{4}(w-x)^2 dx = \left[-\frac{1}{4}(w-x)^3\right]_0^w = \frac{1}{4}w^3$

Case II ($1 < w \le 4$): $P(X+Y \le w) =$

$$= \int_0^{\frac{4(w-1)}{3}} \int_0^{1-\frac{1}{4}x} \frac{3y}{2} dy\, dx + \int_{\frac{4(w-1)}{3}}^{w} \int_0^{w-x} \frac{3y}{2} dy\, dx = \int_0^{\frac{4(w-1)}{3}} \frac{3}{4}(1-\frac{1}{4}x)^2\, dx + \int_{\frac{4(w-1)}{3}}^{w} \frac{3}{4}(w-x)^2 dx =$$

$$= \int_0^{\frac{4(w-1)}{3}} \frac{3}{64}(x-4)^2\, dx + \int_{\frac{4(w-1)}{3}}^{w} \frac{3}{4}(x-w)^2 dx = \left[\frac{1}{64}(x-4)^3\right]_0^{\frac{4(w-1)}{3}} + \left[\frac{1}{4}(x-w)^3\right]_{\frac{4(w-1)}{3}}^{w} =$$

$$= \frac{1}{64}\left(\frac{4(w-1)}{3} - 4\right)^3 + 1 + 0 - \frac{1}{4}\left(\frac{4(w-1)}{3} - w\right)^3 = \left(\frac{w-1}{3} - 1\right)^3 + 1 - \frac{1}{4\times 27}\left(4(w-1) - 3w\right)^3$$

$$= \frac{1}{27}\left(w-4\right)^3 + 1 - \frac{1}{4\times 27}\left(w-4\right)^3 = \frac{1}{36}\left(w-4\right)^3 + 1$$

After taking the derivatives, we obtain the pdf:

$$f_W(w) = \begin{cases} \frac{3}{4}w^2 & \text{for } w \in (0\,,\,1] \\ \frac{1}{12}w^2 - \frac{2}{3}w + \frac{4}{3} & \text{for } w \in (1\,,\,4] \end{cases}$$

◄

6.1.2 *The transformation method*

(B&E, pages 197-209)

When applying the transformation method, we will, just as we did for the univariate case, start with the situation where we are dealing with *discrete* random variables.

Example 6.6

If X_1 and X_2 are two independent discrete random variables, each GEO(p) distributed, then from Theorem 5.9 it follows that:

$f_{X_1,X_2}(x_1,x_2) = f_{X_1}(x_1) \cdot f_{X_2}(x_2) = pq^{x_1-1}pq^{x_2-1} = p^2q^{x_1+x_2-2}$ for $x_1 = 1,2,\ldots$ and $x_2 = 1,2,\ldots$

Then what is the joint pdf of $Y_1 = X_1$ and $Y_2 = X_1 + X_2$?

$$f_{Y_1,Y_2}(y_1,y_2) = P(Y_1 = y_1, Y_2 = y_2) = P(X_1 = y_1, X_1 + X_2 = y_2)$$

$$= P(X_1 = y_1, X_2 = y_2 - y_1)$$
$$= f_{X_1,X_2}(y_1, y_2 - y_1) = p^2q^{y_2-2} \quad \text{for } y_1 = 1,2,\ldots \text{ and } y_2 = y_1+1, y_1+2\ldots$$ ◄

In the example above, a 1-to-1 transformation was defined of the 2-dimensional random variable (X_1,X_2) to (Y_1,Y_2) via the 2-dimensional function $\boldsymbol{u}(x_1,x_2) = (u_1(x_1,x_2),\, u_2(x_1,x_2))$, where $Y_1 = u_1(X_1,X_2) = X_1$ and $Y_2 = u_2(X_1,X_2) = X_1 + X_2$, or, in vector notation: $\boldsymbol{Y} = \boldsymbol{u}(\boldsymbol{X})$. We can solve the set of equations $y_1 = u_1(x_1,x_2) = x_1$ and $y_2 = u_2(x_1,x_2) = x_1 + x_2$: the solution x_1, x_2 expresses x_1 and x_2 as functions of y_1 and y_2, in this case $x_1 = y_1$ and $x_2 = y_2 - y_1$. This gives rise to the following theorem:

Theorem 6.1 (*B&E, Th.6.3.5*)

If $\boldsymbol{X}$ is a k-dimensional discrete random variable with joint pdf $f_{\boldsymbol{X}}(\boldsymbol{x})$, and $\boldsymbol{Y} = \boldsymbol{u}(\boldsymbol{X})$ defines a 1-to-1 transformation, then the joint pdf of $\boldsymbol{Y}$ is:

$$f_{\boldsymbol{Y}}(y_1, y_2, \ldots, y_k) = f_{\boldsymbol{X}}(x_1, x_2, \ldots, x_k)$$

where $x_1, x_2, \ldots, x_k$ are the solutions to the equations $\boldsymbol{y} = \boldsymbol{u}(\boldsymbol{x})$.

Although this theorem might look unnecessarily complicated, this is exactly what happened in Example 6.6. If the transformation is not 1-to-1, then it is still often possible to split the support set of $\boldsymbol{X}$ into disjoint subsets A_j, such that the transformation is 1-to-1 on each of the subsets. In order to determine the joint pdf of $\boldsymbol{Y}$, we should then sum over all these subsets (compare this with Example 4.22 of the previous course for a univariate example, and/or see B&E, Eq. 6.3.16).

For a *continuous* random variable X with the 1-to-1 transformation $Y = u(X)$, we derived in the previous course that the pdf of Y can be written as:

$$f_Y(y) = f_X(u^{-1}(y))\left|\frac{d}{dy}u^{-1}(y)\right|$$

Here, $x = u^{-1}(y)$ denotes the solution to the equation $y = u(x)$.

Example 6.7

In Theorem 4.15, we used the CDF-method to prove that if $X \sim \mathrm{GAM}(\theta, r)$, then Y defined by $Y = 2X/\theta \sim \chi^2(2r)$. We will now use the transformation method to prove the same.

Because $f_X(x) = \dfrac{1}{\theta^r\,\Gamma(r)} x^{r-1} e^{-x/\theta}$ for $x \geq 0$,

and since $x = u^{-1}(y) = \theta y/2$, it follows that

$$f_Y(y) = f_X(\theta y/2)\frac{\theta}{2} = \frac{1}{\theta^r\,\Gamma(r)2^{r-1}}\theta^{r-1} y^{r-1} e^{-y/2}\frac{\theta}{2} = \frac{1}{\Gamma(r)2^r} y^{r-1} e^{-y/2} \qquad \text{for } y \geq 0$$

which is indeed the pdf of $Y \sim \chi^2(2r)$. ◄

In the k-dimensional case with the transformation $\boldsymbol{Y} = \boldsymbol{u}(\boldsymbol{X})$, where $x_1, x_2, ..., x_k$ are the solutions to the equations $\boldsymbol{y} = \boldsymbol{u}(\boldsymbol{x})$ (so of $y_i = u_i(x_1, ..., x_k)$ for $i = 1, \dots, k$), we will define the **Jacobian J** as the determinant of the matrix with all first-order partial derivatives of x_i with respect to y_j:

$$J = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} & \cdots & \frac{\partial x_1}{\partial y_k} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} & \cdots & \frac{\partial x_2}{\partial y_k} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial x_k}{\partial y_1} & \frac{\partial x_k}{\partial y_2} & \cdots & \frac{\partial x_k}{\partial y_k} \end{vmatrix}$$

Remark. The term 'Jacobian' is in the literature used for both the matrix of first-order partial derivatives as for the determinant of that matrix. Jacobians are important in Calculus in situations where one set of coordinates is mapped onto another set (see also Appendix A.6). Furthermore, numerical techniques for finding the solution to a set of non-linear equations in an iterative way often rely heavily on the use of Jacobian matrices.

The following theorem is stated without proof:

Theorem 6.2 *(B&E, Th.6.3.6)*

If $\boldsymbol{X}$ is a k-dimensional continuous random variable with joint pdf $f_{\boldsymbol{X}}(\boldsymbol{x})$ on the support set A, and $\boldsymbol{Y} = \boldsymbol{u}(\boldsymbol{X})$ defines a 1-to-1 transformation of A to the range of A, then the joint pdf of $\boldsymbol{Y}$ is given by:

$$f_{\boldsymbol{Y}}(y_1, y_2, ..., y_k) = f_{\boldsymbol{X}}(x_1, x_2, ..., x_k)\,|\,J\,|$$

(as long as the Jacobian is continuous and non-zero over the range of the transformation), where $x_1, x_2, ..., x_k$ are the solutions (in terms of $y_1, y_2, ..., y_k$) of the equations $\boldsymbol{y} = \boldsymbol{u}(\boldsymbol{x})$.

Example 6.8

Let X_1 and X_2 be independent random variables with $X_1 \sim \text{EXP}(1)$ en $X_2 \sim \text{EXP}(1)$. Thus:

$$f_{X_1,X_2}(x_1,x_2) = f_{X_1}(x_1) f_{X_2}(x_2) = e^{-x_1} \cdot e^{-x_2} = e^{-(x_1+x_2)} \qquad x_1 > 0, x_2 > 0$$

We will now determine the joint pdf of $Y_1 = X_1$ and $Y_2 = X_1 + X_2$. The transformation $y_1 = x_1$ and $y_2 = x_1 + x_2$ has the solution $x_1 = y_1$ and $x_2 = y_2 - y_1$ (unique solution, so the transformation is 1-to-1; also see Example 6.6). The Jacobian is therefore

$$J = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{vmatrix} = \begin{vmatrix} 1 & 0 \\ -1 & 1 \end{vmatrix} = 1$$

$$\Rightarrow \quad f_{Y_1,Y_2}(y_1,y_2) = f_{X_1,X_2}(y_1, y_2 - y_1) \cdot 1 = e^{-y_2}$$

We still need to find the support of (Y_1, Y_2). This can be found by determining the image of the support of (X_1, X_2) by applying the transformation. In this case is it clear that $y_1 > 0$ and $y_2 > y_1$, or: $0 < y_1 < y_2$.

Of course, we can now also find the marginal pdf's of Y_1 and Y_2 by integration over the joint pdf.◄

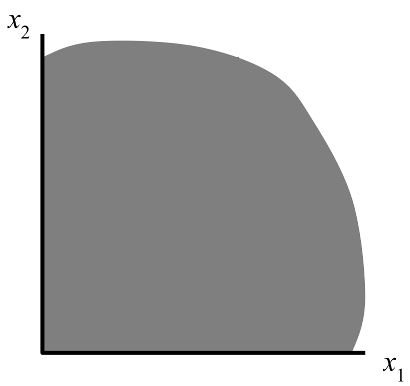
Example 6.9

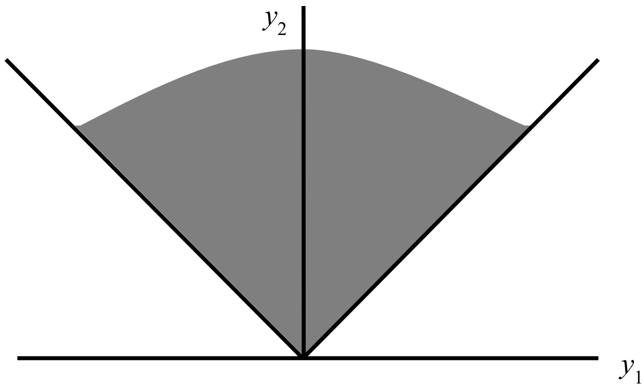
Again consider the same random variables X_1 and X_2 as in the previous example, but now we will look at the transformation $y_1 = x_1 - x_2$ and $y_2 = x_1 + x_2$. The solution to this transformation is $x_1 = (y_1 + y_2)/2$ and $x_2 = (y_2 - y_1)/2$. The Jacobian is therefore

$$J = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{vmatrix} = \begin{vmatrix} 1/2 & 1/2 \\ -1/2 & 1/2 \end{vmatrix} = 1/2$$

$$\Rightarrow \quad f_{Y_1,Y_2}(y_1,y_2) = f_{X_1,X_2}((y_1 + y_2)/2,\ (y_2 - y_1)/2) \cdot 1/2$$
$$= \tfrac{1}{2} e^{-(y_1+y_2)/2} e^{-(y_2-y_1)/2} = \tfrac{1}{2} e^{-y_2} \qquad \text{for } y_2 > 0,\ -y_2 < y_1 < y_2$$

One way of finding the proper support, is to replace x_1 and x_2 in the support of (X_1, X_2) by the solutions $x_1 = (y_1 + y_2)/2$ and $x_2 = (y_2 - y_1)/2$. We obtain $(y_1 + y_2)/2 > 0$ and $(y_2 - y_1)/2 > 0$. These two inequalities can be then rewritten as: $y_2 > 0$ and $-y_2 < y_1 < y_2$ (Check!)





Drawing the support sets can also be helpful in finding the support of (Y_1, Y_2) (see figures). Usually, it is enough to determine how the borders of the support of (X_1, X_2) are transformed; in this case we can see that the line $x_1 = 0$ is transformed into the line $y_2 = -y_1$), and the line $x_2 = 0$ is transformed into the line $y_2 = y_1$):

Remark. Often the support set can be formulated in different, but equivalent, ways. ◄

Because this transformation method can be applied only when the transformation is one-to-one, we can apply it only to a *k*-dimensional random variable if the transformation itself is a *k*-dimensional function. However, this method can still be useful, even if the sole purpose lies in finding the distribution of a single function of *k* random variables, as illustrated below.

Example 6.10

Let X_1 and X_2 be two independent random variables, both uniformly distributed on (0, 1), such that $f_{X_1,X_2}(x_1,x_2) = 1$ for $0 \le x_1, x_2 \le 1$. Assume we would like to find the distribution of $Y_1 = X_1 X_2$. Since this is clearly not a 1-to-1 transformation, we need first to define a second conveniently chosen function, like for example $Y_2 = X_1$. Now, we do have a 1-to-1 transformation of (X_1, X_2) onto (Y_1, Y_2), with the solutions $x_1 = y_2$ and $x_2 = y_1 / y_2$. The Jacobian is then

$$J = \begin{vmatrix} \dfrac{\partial x_1}{\partial y_1} & \dfrac{\partial x_1}{\partial y_2} \\ \dfrac{\partial x_2}{\partial y_1} & \dfrac{\partial x_2}{\partial y_2} \end{vmatrix} = \begin{vmatrix} 0 & 1 \\ 1/y_2 & -y_1/y_2^2 \end{vmatrix} = -1/y_2$$

Now the joint pdf of (Y_1, Y_2) follows:

$$\Rightarrow \quad f_{Y_1,Y_2}(y_1,y_2) = f_{X_1,X_2}(y_2,\ y_1/y_2) \cdot \left|-1/y_2\right|$$
$$= 1 \cdot 1/y_2 = \frac{1}{y_2} \quad \text{for } 0 \le y_1 \le y_2 \le 1$$

But, since we are not interested in this joint pdf, but only in the (marginal) pdf of Y_1, we get:

$$f_{Y_1}(y_1) = \int_{y_1}^{1} \frac{1}{y_2} dy_2 = -\ln y_1 \quad \text{for } 0 \le y_1 \le 1$$

Note that this result is (of course) the same as what we found in Example 6.3. ◄

Example 6.11

Let X_1, X_2 and X_3 be independent random variables, with $X_i \sim \text{GAM}(1,\ \alpha_i),\ i = 1,2,3$, so

$$f_{X_1,X_2,X_3}(x_1,x_2,x_3) = \prod_{i=1}^{3} \frac{1}{\Gamma(\alpha_i)} x_i^{\alpha_i - 1} e^{-x_i} \qquad 0 < x_i$$

Consider the transformation $Y_1 = X_1/(X_1 + X_2 + X_3)$, $Y_2 = X_2/(X_1 + X_2 + X_3)$ and $Y_3 = X_1 + X_2 + X_3$, with as inverse transformation $x_1 = y_1 y_3$, $x_2 = y_2 y_3$ and $x_3 = y_3(1 - y_1 - y_2)$. The Jacobian is:

$$J = \begin{vmatrix} y_3 & 0 & y_1 \\ 0 & y_3 & y_2 \\ -y_3 & -y_3 & 1 - y_1 - y_2 \end{vmatrix} = y_3(y_3(1 - y_1 - y_2) + y_2 y_3) - y_3(-y_1 y_3) = y_3^2$$

so

$$f_{Y_1,Y_2,Y_3}(y_1,y_2,y_3) = \frac{y_1^{\alpha_1 - 1} y_2^{\alpha_2 - 1} (1 - y_1 - y_2)^{\alpha_3 - 1}\ y_3^{\alpha_1 + \alpha_2 + \alpha_3 - 1} e^{-y_3}}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)} \qquad 0 < y_i \text{ and } y_1 + y_2 < 1$$

Assuming we need to find the pdf of Y_3, note that we could do so by means of integration. However, we will use as an illustration another way of finding this pdf which is possible in this particular case. Note the joint pdf above can be factorised in a product of a function of $g(y_3)$ of y_3 and a function of both y_1 and y_2 (choose for example $g(y_3) = y_3^{\alpha_1+\alpha_2+\alpha_3-1}e^{-y_3}$). Also, the support of (Y_1,Y_2,Y_3) is clearly a Cartesian product of the supports of Y_3 and (Y_1,Y_2). This means (analogous to Theorem 5.10): Y_3 and (Y_1,Y_2) are independent of each other (note however, that Y_1 and Y_2 are *not* independent). This also means that the pdf of Y_3 is equal to the $g(y_3)$ multiplied by a certain constant *c* such that the total area under its curve is equal to 1:

$$f_{Y_3}(y_3) = cg(y_3) = cy_3^{\alpha_1+\alpha_2+\alpha_3-1}e^{-y_3} \qquad 0 < y_3$$

From the form of this pdf, we can (even without having to determine the value of *c* first) recognise that $Y_3 = X_1 + X_2 + X_3 \sim$ GAM(1, $\alpha_1+\alpha_2+\alpha_3$).

(**Remark.** This result could have been found in a much simpler way using mgf's) ◄

Remark. If a transformation of the *k*-dimensional space to another *k*-dimensional space is not 1-to-1 on the support of ***X***, we should try to split the support set into disjoint subsets A_j, such that the transformation is 1-to-1 on each of the separate subsets. In order to find the joint pdf of ***Y***, the formula in Theorem 6.2 should be summed over all these subsets (see e.g. B&E, Eq. 6.3.20). We will not bother about those situations here anymore.

6.1.3 The moment generating function method

In § 4.6.3 this method was discussed for determining the distribution of a univariate random variable. As a result, we can easily prove that the sum of two independent normal random variables *X* and *Y* is itself also (univariate) normally distributed with expected value $\mu_X + \mu_Y$ and variance $\sigma_X^2 + \sigma_Y^2$. A similar way can sometimes be helpful as well for finding joint distributions, using the joint mgf (see § 5.6). (Bain and Engelhardt does not give examples of such an application of this method).

Example 6.12

Let *X* and *Y* be two independent standard normal random variables. So we know:

$$M_X(t) = e^{t^2/2},\ M_Y(t) = e^{t^2/2}$$

Consider the transformation *U* = *X* + *Y* and *W* = *X* – *Y*. Because *X* and *Y* are independent, it follows immediately that $M_{X+Y}(t) = M_X(t)M_Y(t) = e^{t^2}$, which shows that *U* ~ N(0, 2).

Because of the symmetry of the standard normal distribution, it follows that $M_{-Y}(t) = e^{t^2/2}$, and thus that $M_{X-Y}(t) = M_X(t)M_{-Y}(t) = e^{t^2}$.

Now we will attempt to find the joint distribution of *U* and *W* :

$$\begin{aligned} M_{U,W}(t_1,t_2) &= \mathrm{E}[e^{t_1U+t_2W}] = \mathrm{E}[e^{t_1(X+Y)+t_2(X-Y)}] \\ &= \mathrm{E}[e^{(t_1+t_2)X+(t_1-t_2)Y}] \\ &= \mathrm{E}[e^{(t_1+t_2)X}]\mathrm{E}[e^{(t_1-t_2)Y}] \qquad \text{(because of Theorem 5.11 (1a))} \\ &= M_X(t_1+t_2)\cdot M_Y(t_1-t_2) \\ &= e^{(t_1+t_2)^2/2}\cdot e^{(t_1-t_2)^2/2} \\ &= e^{(2t_1^2+2t_2^2)/2} \end{aligned}$$

This joint mgf can be recognised as the mgf of a bivariate normal distribution: $(U,W) \sim \mathrm{BVN}(0,0,2,2,0)$ (check!). The value of ρ (=0) might seem surprising, since it means that *U* and *W* are uncorrelated and even independent!

Note also that this joint mgf can be used to find immediately the mgf's of *U* = *X* + *Y* and *W* = *X* – *Y* by substituting $t_2 = 0$ and $t_1 = 0$ respectively. ◄

6.2 Distributions of sums of random variables

Often, we are interested to know the distribution of sums of random variables, so of $\sum_{i=1}^{k} X_i$. We have seen already many of these results of which we will list the most important ones again here.

Special attention is paid to those sums where the random variables are not only independent, but also identically distributed, i.e. $f_{X_1}(x) = f_{X_2}(x) = \cdots = f_{X_k}(x)$. In these situations, the subscript *i* can often be deleted; we write $f_X(x)$, $\mathrm{E}(X)$, μ_X and $M_X(t)$ instead of $f_{X_i}(x)$, $\mathrm{E}(X_i)$, μ_{X_i} and $M_{X_i}(t)$ respectively. But note: we cannot always simply delete all subscripts. For example: even though X_i and X_j (for all $i \neq j$) are independent random variables, the same can definitely not be said when we delete the subscripts, since *X* and *X* are *not* independent! In the literature, random variables which are both mutually independent as well as identically distributed are often indicated by the abbreviation 'i.i.d.' (Independent and identically distributed).

Remark. Bain&Engelhardt do not use the abbreviation i.i.d., but use the term 'random sample' with which they mean exactly the same (see also 6.4.1).

6.2.1 Expected values and variances of sums of random variables

Always: $\mathrm{E}(\sum_{i=1}^{k} X_i) = \sum_{i=1}^{k} \mathrm{E}(X_i)$ (Theorem 5.7)

→ If identically distributed: $\mathrm{E}(\sum_{i=1}^{k} X_i) = k\ \mathrm{E}(X)$

Always: $\mathrm{Var}(\sum_{i=1}^{k} X_i) = \sum_{i=1}^{k} \mathrm{Var}(X_i) + 2\sum\sum_{i<j} \mathrm{Cov}(X_i, X_j)$ (Theorem 5.18)

→ If independent: $\mathrm{Var}(\sum_{i=1}^{k} X_i) = \sum_{i=1}^{k} \mathrm{Var}(X_i)$ (also Theorem 5.18)

→ If i.i.d.: $\mathrm{Var}(\sum_{i=1}^{k} X_i) = k\,\mathrm{Var}(X)$

6.2.2 Distributions of sums of independent random variables

Below, we provide a list with the distribution of $Y = \sum_{i=1}^{k} X_i$ in a number of specific cases, where each time the random variables are mutually independent. These results can easily be derived by using moment generating functions, since we know from Theorem 5.17 that if the random variables $X_1, \ldots, X_k$ are mutually independent, then

$$M_Y(t) = M_{X_1+\ldots+X_k}(t) = \prod_{i=1}^{k} M_{X_i}(t).$$

If X_1 to X_k are also identically distributed, then

$$M_Y(t) = M_{X_1+\ldots+X_k}(t) = \left(M_X(t)\right)^k.$$

Thus, we obtain:

If $X_i \sim \mathrm{BIN}(m_i, p)$, then $Y \sim \mathrm{BIN}(\sum_{i=1}^{k} m_i, p)$

If $X_i \sim \mathrm{BIN}(m, p)$, then $Y \sim \mathrm{BIN}(km, p)$

If $X_i \sim \mathrm{POI}(\mu_i)$, then $Y \sim \mathrm{POI}(\sum_{i=1}^{k} \mu_i)$

If $X_i \sim \text{POI}(\mu)$, then $Y \sim \text{POI}(k\mu)$

If $X_i \sim \text{GEO}(p)$, then $Y \sim \text{NEGBIN}(k, p)$

If $X_i \sim \text{NEGBIN}(r_i, p)$, then $Y \sim \text{NEGBIN}(\sum_{i=1}^{k} r_i, p)$

If $X_i \sim \text{NEGBIN}(r, p)$, then $Y \sim \text{NEGBIN}(kr, p)$

If $X_i \sim \text{EXP}(\lambda)$, then $Y \sim \text{GAM}(\frac{1}{\lambda}, k)$

If $X_i \sim \text{GAM}(\theta, r_i)$, then $Y \sim \text{GAM}(\theta, \sum_{i=1}^{k} r_i)$

If $X_i \sim \text{GAM}(\theta, r)$, then $Y \sim \text{GAM}(\theta, kr)$

If $X_i \sim \text{N}(\mu_i, \sigma_i^2)$, then $Y \sim \text{N}(\sum_{i=1}^{k} \mu_i, \sum_{i=1}^{k} \sigma_i^2)$

If $X_i \sim \text{N}(\mu, \sigma^2)$, then $Y \sim \text{N}(k\mu, k\sigma^2)$

If $X_i \sim \chi^2(\nu_i)$, then $Y \sim \chi^2(\sum_{i=1}^{k} \nu_i)$

If $X_i \sim \chi^2(\nu)$, then $Y \sim \chi^2(k\nu)$

Remark. Although we can see from this list that the distribution of the sum often has a similar type of distribution as the distribution of the constituting random variables, it should be emphasised here that this is not always the case! For example: if $X_i \sim \text{BIN}(m, p_i)$, then the distribution of Y is *not* binomially distributed as well (unless $p_1 = p_2 = ... = p_n$).

6.2.3 Convolution formula

The convolution formula below can be helpful in the determination of the sum $S = X + Y$ of two independent random variables. In section 3.6 of the previous course, the convolution formula has been given for the sum $S = X + Y$ of two *discrete and independent random variables which can both only attain values on the set of the natural numbers.* This formula was:

$$P(S = s) = \sum_{i=0}^{s} P(X = i)\, P(Y = s - i) \quad \text{or} \quad f_S(s) = \sum_{i=0}^{s} f_X(i)\, f_Y(s - i)$$

For continuous random variables a similar result can be derived:

Theorem 6.3 (*B&E, Eq.6.4.2*)

If X and Y are two independent continuous random variables, then the pdf of $S = X + Y$ is:

$$f_S(s) = \int_{-\infty}^{\infty} f_X(t)\, f_Y(s - t) dt$$

This is the **convolution formula**.

Proof

We choose here to use the transformation method in order to prove this theorem. Define $S = X + Y$ and $T = X$, resulting in the inverse transformation: $x = t$ and $y = s - t$. The Jacobian is therefore

$$J = \begin{vmatrix} 1 & 0 \\ -1 & 1 \end{vmatrix} = 1$$

$$\Rightarrow \quad f_{T,S}(t, s) = f_{X,Y}(t,\ s - t)$$

$$\Rightarrow \quad f_S(s) = \int_{-\infty}^{\infty} f_{T,S}(t,\ s) dt = \int_{-\infty}^{\infty} f_{X,Y}(t,\ s - t) dt$$

The result then follows after using the information that X and Y are independent.

Example 6.13

Let X and Y be two independent random variables, both EXP(λ) distributed.
The pdf of the sum $S = X + Y$ is:

$$f_S(s) = \int_{-\infty}^{\infty} f_X(t)\, f_Y(s-t)dt = \int_0^s \lambda e^{-\lambda t}\ \lambda e^{-\lambda(s-t)}dt = \int_0^s \lambda^2 e^{-\lambda s}dt = \lambda^2 s e^{-\lambda s} \quad \text{for } 0 < s$$

(Compare this with Example 6.2). ◄

6.3 The t-distribution and the F-distribution.

(B&E, pages 273-277)

The t-distribution

Definition 6.1

A continuous random variable T has a ***t*-distribution** with parameter ν (notation: $T \sim t_\nu$ or $T \sim t(\nu)$) when the pdf is given by:

$$f(x) = \frac{\Gamma(\frac{1}{2}\nu + \frac{1}{2})}{\Gamma(\frac{1}{2}\nu)\sqrt{\pi\nu}}(1 + \frac{x^2}{\nu})^{-\frac{1}{2}(\nu+1)} \qquad (-\infty < x < \infty)$$

where ν = 1, 2, 3, … . This parameter is usually called the 'number of degrees of freedom' (Dutch: 'aantal vrijheidsgraden') (Note that $\Gamma(\cdot)$ represents, as before, the Gamma-function, see previous course and the Appendix).

The next theorem in combination with Theorem 7.1 will make clear why this *t*-distribution is so important within the field of inferential statistics (see chapters 7 and 8).

Theorem 6.4 (*B&E, Th.8.4.1*)

If $Z \sim N(0, 1)$, $U \sim \chi^2(\nu)$, and Z and U are independent, then:

$$T = \frac{Z}{\sqrt{U/\nu}} \sim t(\nu)$$

(so T has a *t*-distribution with *v* degrees of freedom.

Proof

(It is enough to study only the logic and the structure of this proof). Because it is given that Z and U are independent, the joint pdf can be written as:

$$f_{Z,U}(z,u) = f_Z(z) f_U(u) = \frac{u^{\nu/2-1}e^{-u/2}e^{-z^2/2}}{\sqrt{2\pi}\Gamma(\nu/2)2^{\nu/2}} \quad \text{for } 0 < u$$

We consider now the transformation $T = \dfrac{Z}{\sqrt{U/\nu}}$. Before we can apply the transformation method, we first need to define a second random variable such that a 1-to-1 transformation results. We will do this in such a way that the inverse transformation is rather straightforward. Here we use $W = U$, with the inverse transformation $u = w$ and $z = t\sqrt{w/\nu}$. The Jacobian is then $\sqrt{w/\nu}$, so we obtain

$$f_{T,W}(t,w) = \frac{w^{\nu/2-1}e^{-w/2}e^{-t^2 w/2\nu}(w/\nu)^{1/2}}{\sqrt{2\pi}\Gamma(\nu/2)2^{\nu/2}} \quad \text{for } 0 < w$$

The pdf of T results by integration of the joint pdf with respect to w:

$$\Rightarrow \quad f_T(t) = \int_0^\infty \frac{w^{\nu/2-1} e^{-w/2} e^{-t^2 w/2\nu} (w/\nu)^{1/2}}{\sqrt{2\pi}\Gamma(\nu/2)2^{\nu/2}} dw$$

$$= \frac{1}{\sqrt{2\pi\nu}\Gamma(\nu/2)2^{\nu/2}} \int_0^\infty w^{(\nu+1)/2-1} e^{-w(\nu+t^2)/2\nu} dw$$

$$= \frac{\Gamma((\nu+1)/2)}{\sqrt{2\pi\nu}\Gamma(\nu/2)2^{\nu/2}} \left(\frac{2\nu}{\nu+t^2}\right)^{(\nu+1)/2} \int_0^\infty \frac{((\nu+t^2)/2\nu)^{(\nu+1)/2} w^{(\nu+1)/2-1} e^{-w(\nu+t^2)/2\nu}}{\Gamma((\nu+1)/2)} dw$$

$$= \frac{\Gamma((\nu+1)/2)}{\sqrt{\pi\nu}\Gamma(\nu/2)} \left(\frac{\nu}{\nu+t^2}\right)^{(\nu+1)/2}$$

$$= \frac{\Gamma((\nu+1)/2)}{\sqrt{\pi\nu}\Gamma(\nu/2)} \left(\frac{\nu+t^2}{\nu}\right)^{-(\nu+1)/2}$$

(the last integral above is equal to 1, because the integrand is the pdf of a Gamma-distributed random variable, $\text{GAM}(((\nu+t^2)/2\nu)^{-1},\ (\nu+1)/2)$).
The resulting formula is indeed the pdf of a t-distributed random variable with ν degrees of freedom.

The next theorem gives the expected value and the variance of the *t*-distribution:

Theorem 6.5 (≈*B&E, Th.8.4.2*)

If $T \sim t(\nu)$, then for $\nu > 2$:

$$\text{E}(T) = 0 \quad \text{and} \quad \text{Var}(T) = \frac{\nu}{\nu-2}$$

<u>*Proof*</u>

Recall that *T* can be written as $T = \dfrac{Z}{\sqrt{U/\nu}}$ with *Z* and *U* as in Theorem 6.4. Then:

$$\text{E}(T) = \text{E}\left[\frac{Z}{\sqrt{U/\nu}}\right] = \sqrt{\nu}\text{E}(Z)\text{E}(1/\sqrt{U}) \quad \text{(because of the independence of } Z \text{ and } U\text{)}$$

Because E(*Z*) = 0, it follows that E(*T*) = 0. The second moment of *T* is:

$$\text{E}(T^2) = \text{E}\left[\frac{Z^2}{U/\nu}\right] = \nu\text{E}(Z^2)\text{E}(1/U)$$

We know already that $\text{E}(Z^2) = 1$, so we only need to find $\text{E}(1/U)$. We do that here using the pdf of the chi-square distribution:

$$\text{E}\left[\frac{1}{U}\right] = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} \int_0^\infty \frac{1}{u} e^{-u/2} u^{\nu/2-1} du$$

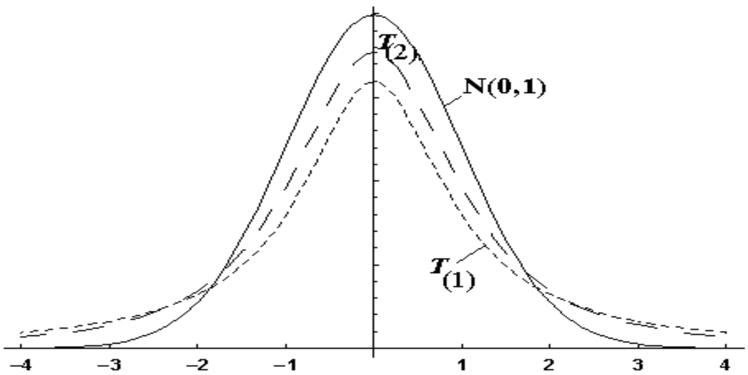
$$= \frac{2^{(\nu-2)/2}\Gamma((\nu-2)/2)}{2^{\nu/2}\Gamma(\nu/2)} \int_0^\infty \frac{e^{-u/2} u^{(\nu-2)/2-1}}{2^{(\nu-2)/2}\Gamma((\nu-2)/2)} du$$

$$= \frac{2^{(\nu-2)/2}\Gamma((\nu-2)/2)}{2^{\nu/2}\Gamma(\nu/2)} = \frac{1}{2(\nu-2)/2} = \frac{1}{\nu-2}$$

Note that the integral above yields 1, since the integrand is the pdf of a $\chi^2(\nu-2)$ random variable (but only if $\nu > 2$!).
Entering this into $\text{E}(T^2) = \nu\text{E}(Z^2)\text{E}(1/U)$ concludes the proof.

Unfortunately, the moment generating function for a *t*-distribution does not exist… (the characteristic function, mentioned briefly in the previous course, does exist, but we will not need it here).

The theorem above shows that the variance of T converges to 1 when the number of degrees of freedom increases to infinity. Also, inspection of the pdf shows that the t-distribution is symmetrical around 0. In fact, its pdf has a shape very similar to the shape of the pdf of the standard normal distribution. The top (at $t = 0$) is somewhat lower, and its tails are thicker. For large values of ν, the *t*-distribution becomes indistinguishable from the standard normal distribution: ***the t-distribution is asymptotically equal to the standard normal distribution.*** Note also that the denominator in T, $\sqrt{U/\nu}$, will converge to 1 when ν goes to infinity (why? Check the expected value and the variance of U; for a formal proof, we need techniques discussed in the course Prob. Theory and Statistics 3).



Table

Tables for 'the' *t*-distribution usually consist of one page with right-tail critical values, with one row for each number of degrees of freedom. Above each column, the probability of exceeding the tabulated value is listed. The table in the Appendix of this reader shows for example that $P(T_3 > 1.638) = 0.1$. The value 1.638 will be written as $t_{0.1;\,3}$ or simply as $t_{0.1}$ if there is no need to stress the number of degrees of freedom. We can also say that $t_{0.1}$ represents the 1–0.1 = 0.9 quantile, or the 90-th percentile of the *t*-distribution.
The last row in the table relates to an infinite number of degrees of freedom. As we have seen, the *t*-distribution is then equal to the standard normal distribution. Thus, if we need to find the value z_α such that $P(Z \geq z_\alpha) = \alpha$, then that is where this value can be found! (B&E use the left-tail critical values in the tables instead, so be aware).

The F-distribution

This distribution is very important as well in inferential statistics, especially within linear regression.

> ***Definition 6.2***
> A continuous random variable F has the ***F*-distribution** with parameters ν_1 and ν_2 (notation: $F \sim F_{\nu_1,\nu_2}$ or $F \sim F(\nu_1,\nu_2)$) when its pdf is given by:
>
> $$f(x) = \frac{\Gamma(\frac{1}{2}\nu_1 + \frac{1}{2}\nu_2)}{\Gamma(\frac{1}{2}\nu_1)\Gamma(\frac{1}{2}\nu_2)} \sqrt{\frac{\nu_1^{\nu_1}\nu_2^{\nu_2}x^{\nu_1-2}}{(\nu_2+\nu_1 x)^{\nu_1+\nu_2}}} \qquad \text{for } x > 0$$
>
> where ν_1 , ν_2 = 1, 2, 3, … .
> The parameter ν_1 is usually called the number of degrees of freedom in the numerator, and ν_2 as the number of degrees of freedom in the denominator.

Often, we will use the same notation for the random variable itself as well, so e.g. $F_{\nu_1,\nu_2} \sim F_{\nu_1,\nu_2}$.

Theorem 6.6 (*B&E, Th.8.4.4*)

If $U_1 \sim \chi^2(\nu_1)$ and $U_2 \sim \chi^2(\nu_2)$ are independent random variables, then:

$$F = \frac{U_1/\nu_1}{U_2/\nu_2} \sim F(\nu_1, \nu_2)$$

<u>*Proof*</u>

Very analogous to the proof of Theorem 6.4.

Theorem 6.7 (*B&E, Th.8.4.5*)

If $F \sim F(\nu_1, \nu_2)$, then for $\nu_1 > 4$:

$$\mathrm{E}(F) = \frac{\nu_2}{\nu_2 - 2} \quad \text{and} \quad \mathrm{Var}(F) = \frac{2\nu_2^2(\nu_1 + \nu_2 - 2)}{\nu_1(\nu_2 - 2)^2(\nu_2 - 4)}$$

<u>*Proof*</u>

Since F can be written as $F = \dfrac{U_1/\nu_1}{U_2/\nu_2}$ with U_1 and U_2 like in Theorem 6.6, we get:

$$\mathrm{E}(F) = \mathrm{E}\left[\frac{U_1/\nu_1}{U_2/\nu_2}\right] = \frac{\nu_2}{\nu_1}\mathrm{E}(U_1)\mathrm{E}(1/U_2) \quad \text{(because of independence of } U_1 \text{ and } U_2\text{)}$$

In Theorem 6.5, we showed already that $\mathrm{E}(1/U_2) = \dfrac{1}{\nu_2 - 2}$.

In the course Prob. Theory and Statistics 1, we saw that $\mathrm{E}(U_1) = \nu_1$.

Substitution in $\mathrm{E}(F)$ gives now the requested result.

The proof for the variance is not given here (you might try this yourself).

Check that from the theorems above, it follows that an F-distributed random variable with ν_1=1 and ν_2=ν has the same distribution as the square of t_ν-distributed random variable!

Table

A table for the F-distribution usually consists of several pages with only right-tail critical values w, $P(F > w) = \alpha$ for small values of α. In the Appendix of this reader, only critical values for $\alpha = 0.05$ are shown; check Canvas for more complete tables.

<u>*Example 6.14*</u>

$P(F_{3,5} > 5.41) = 0.05$ (F-tabel, $\alpha = 0.05$, third column, fifth row) ◄

Since the F-distribution is not symmetrical, it is slightly more complicated to find a left-tail critical value u, such that $P(F \le u) = \alpha$ for small values of α. We need a kind of trick…. First notice that if $F = \dfrac{U_1/\nu_1}{U_2/\nu_2} \sim F_{\nu_1,\nu_2}$, then of course it follows that $\dfrac{1}{F} = \dfrac{U_2/\nu_2}{U_1/\nu_1} \sim F_{\nu_2,\nu_1}$ (note the swap in the number of degrees of freedom!!). Because $P(F \le u)$ is equal to $P(\frac{1}{F} \ge \frac{1}{u})$, we can now use the tables to find the *right*-tail critical value w, such that $P(\frac{1}{F} \ge w) = \alpha$. By equating $\frac{1}{u} = w$, we can conclude that $u = 1/w$.

<u>*Example 6.15*</u>

Find the value of u such that $P(F_{3,5} < u) = 0.05$. Using the F-table ($\alpha = 0.05$, fifth column, third row) we can see that $P(F_{5,3} > 9.01) = 0.05$. Thus: $u = 1/9.01 = 0.111$, so $P(F_{3,5} < 0.111) = 0.05$. ◄

6.4 Samples

We have been dealing until now mainly with Probability Theory, enabling us to find probabilities for certain events (=outcomes of observations), based on knowledge of the (distribution of the) population from which the observations are taken. In chapter 7, we start the discussion of the field of *inferential statistics,* where we use observations (= samples) to draw conclusions about a population of which the distribution is essentially not (completely) known yet.

In order to do so, we discuss in this section the concept of a sample as a multivariate random variable. We will define some of the most common sample statistics, and we will derive the distributions of a number of important (functions of) sample statistics. Thus, this section acts as the link between probability theory and inferential statistics.

6.4.1 Samples from a distribution

(B&E, pages 158-161, 264-273)

A lot of research is based on samples drawn from a certain population. An observation X in a sample can represent for example the lifespan of a specific light bulb which is part of a population of bulbs. When the distribution of lifespans of all bulbs in the population can be described by the pdf $f(x)$ then we can say that X (when randomly selected from the population) will have a lifespan distribution equal to this pdf, in other words $f_X(x) = f(x)$. When we select more bulbs from this population, we can write X_i as the lifespan of the i-th bulb in this sample. We speak of a *sample from a distribution* if, for each i the distribution of X_i remains the same (so $f_{X_i}(x) = f(x)$) and if furthermore the X_i's are all mutually independent (which means that the outcome of one observations does not influence the outcome of any other observation).

This gives rise to the following definition:

Definition 6.3 (*B&E, Def. 4.6.1*)

The random variables $X_1,...,X_n$ form a **sample from a distribution** (Dutch: steekproef uit een verdeling) with sample size *n,* if the joint pdf (called the **pdf of the sample**) can be written as

$$f_{X_1,...,X_n}(x_1,...,x_n) = f(x_1)\cdot f(x_2)\cdots f(x_n)$$

where $f(x)$ is the pdf of the population distribution.

When we consider a random sample (where each element in the population is equally likely to be selected) from an infinite population, then this results in a sample from a distribution, for example when we throw 10 times with a fair die. However, if the population consists of a finite number of elements, we must distinguish between sampling *with* and *without* replacement. In random **sampling with replacement** each time an element is observed, it will be returned to the population before the next element is drawn. The result will be a row of observations, all representing elements from the population, where it is possible that the same element will be observed more than once. This again represents a sample from a population, since the distribution of the possible outcomes will remain the same at each draw.

But a random **sample without replacement** from a finite population cannot be regarded as a sample from a distribution, because the outcome of the first sampled item will in general change the distribution of the second item, so $X_1,...,X_n$ are not mutually independent. So there is a difference between 'random sample' and 'sample from a population', because it is also possible to speak about a random sample from a finite population. Note however that if the size of the sample is very small compared to the size of the population, we might still treat this situation as if we are dealing with a sample from a distribution, since the degree of dependency will be very limited.

Remark. The book of Bain and Engelhardt defines 'random sample' as meaning exactly the same as what we defined as a 'sample from a distribution'. The terminology in B&E is confusing, since B&E also mentions a ‘random sample from a finite population without replacement’, which does not satisfy their own definition of a ‘random sample’. Unfortunately, this confusion is quite widespread in the literature.)

Example 6.16

Throwing a fair die 600 times and recording all outcomes results in a sample from a distribution (discrete uniform on {1, 2, 3, 4, 5, 6}). Similarly, a sample with replacement of size 600 from a population of the six numbers 1 to 6 is also a sample from (the same) distribution. ◄

The next example illustrates the fact that, given the pdf of a sample, we can find in principle the distribution of any function of $X_1,...,X_n$.

Example 6.17

Again we throw a fair die 600 times, but assume now that we are only interested in number of 6s. Therefore, we define X_i as 1 if the i-th throw results in 6 dots, and 0 otherwise. It will be clear that $X_i \sim \text{BIN}(1, p = \text{"Probability of throwing six dots"})$, with

$$f_{X_i}(x_i) = p^{x_i} q^{1-x_i} \quad (x_i = 0,1).$$

The joint pdf of $(X_1,...,X_{600})$ is:

$$f_{X_1,...,X_{600}}(x_1,\dots,x_{600}) = f(x_1)\cdot f(x_2)\cdots f(x_{600})$$
$$= \prod_{i=1}^{600} p^{x_i} q^{1-x_i} = p^{\sum_{i=1}^{600} x_i} q^{600-\sum_{i=1}^{600} x_i} \text{ (for } x_i = 0,1)$$

This pdf of the sample is a function of 600 variables. But if we are not interested in the order at which the 6s were thrown, but only in the total number of 6s, then we could define the random variable Y as: $Y = \sum_{i=1}^{600} X_i$. The theory of the previous chapter tells us that we can use the joint pdf to find the pdf of Y, so in order to find $P(Y = y)$, we will sum all combinations of x_1 to x_{600} such that $\sum_{i=1}^{600} x_i = y$. Since each of those combinations will appear with the same probability $p^y q^{600-y}$, we can multiply this probability by the number of combinations, which is equal to $\binom{600}{y}$. Thus we obtain:

$$P(Y = y) = \binom{600}{y} p^y q^{600-y} \quad \text{for } y = 0,1,2,...,600 \text{, so } Y \sim \text{BIN}(600, p).$$

In other words: if we are only interested in the number of 6s, then taking a sample of size 600 from the Bernoulli-distribution can be considered to be equal to observing one single draw from a binomial distribution with $n = 600$ and $p = 1/6$.

Of course, there is no need to follow this rather complicated procedure each time, since we have derived this result a few times before already, without using explicitly the joint pdf of the sample. However, in other situations, we will only be able to derive a result by starting with the joint pdf of the sample (see e.g. Theorem 6.14). ◄

Remark. The term ‘random’ defines the procedure followed in the sampling process: there should be no preference what so ever for the selection of some item(s) in the population. Each subset of n elements from the population should have the same probability of being selected as any other subset of the same size. Note that it is impossible to truly judge after drawing a sample whether this was actually done in a random way. For example, when we take a random sample of 10 men from the population of all adult Dutch men, it is still possible that by pure chance it will include four men all taller than 2 meters. So, even though the result might seem not random, it could still represent a true random sample. However, a result like that may be a reason to have another thorough look at the way the men were selected. In such a way, we might discover ‘bad’ samples, but we can never be 100% sure.

6.4.2 *Empirical CDF*

When we consider a sample from a distribution which has the CDF $F(x) = P(X \le x)$, we can define an 'empirical' (CDF) $F_n(x)$. We simply count for each value of x the number of observations in the sample smaller than or equal to x, so:

$$F_n(x) = \frac{\#(x_i \le x)}{n}, \qquad i = 1,\ldots,n$$

An alternative, but equivalent, way is given by the next procedure. We first order all observed values in the sample $x_1,\ldots,x_n$ in increasing magnitude, thus obtaining: $y_1 < y_2 < \cdots < y_n$ with $y_1 = \min(x_1,\ldots,x_n)$ and $y_n = \max(x_1,\ldots,x_n)$. Then it is simple to see that:

$$F_n(x) = \begin{cases} 0 & x < y_1 \\ i/n & y_i \le x < y_{i+1} \\ 1 & y_n \le x \end{cases}$$

This empirical CDF can be displayed in a figure as a step function (see also page 43 of the previous reader). When the sample size n becomes larger and larger, the empirical CDF will look more and more like the population CDF $F(x)$.

Remark. The maximum deviation between these two CDF's can actually be used (in the so-called goodness of fit-tests) to test whether it is possible that a certain sample is drawn from a certain distribution.

6.4.3 *Sample statistics*

Definition 6.4 (*B&E, Def. 8.2.1*)
A function of observable random variables $X_1,\ldots,X_n$, so $T = t(X_1,\ldots,X_n)$, which do not involve any unknown parameters, is called a (**sample**) **statistic** (Dutch: steekproefgrootheid) .

The word 'observable' is essential here, since we will later use statistics to draw conclusions about a population. We can only do so if we can truly observe the values of the random variables $X_1,\ldots,X_n$ in a sample. For the same reason, the definition above also states that T does not involve unknown parameters. Say a population is characterised by an (unknown) population mean μ. Then $X_1 + X_2 + X_3$ is an example of a statistic, but $(X_1 + X_2 + X_3)/\mu$ is not. Note: even though $X_1 + X_2 + X_3$ is a function of $X_1,\ldots,X_n$ which does not involve μ, the *distribution* of $X_1 + X_2 + X_3$ will in general depend on μ!

Numerous possible statistics exist; we will discuss below only some of the most frequently used statistics. Each time, we will determine its expected value and often the variance as well.

Definition 6.5 (*B&E, Eq. 8.2.1*)
Let $X_1,\ldots,X_n$ be a sample. The **sample mean** (Dutch: **s**teekproefgemiddelde) is defined as

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \,.$$

In next theorem will show that the expected value of the sample mean is equal to the population mean, and that the variance of the sample mean becomes smaller when the sample size becomes larger (as long as the population variance exists). These results will play an important role when we want to estimate an unknown population mean by using an observed sample mean (see Chapter 7).

Theorem 6.8 (*B&E, Th.8.2.1*)

If $X_1,...,X_n$ is a sample from a distribution with $\mathrm{E}(X)=\mu$ and $\mathrm{Var}(X)=\sigma^2$, then:

$$\mathrm{E}(\bar{X})=\mu \quad \text{and} \quad \mathrm{Var}(\bar{X})=\sigma_{\bar{X}}^2=\frac{\sigma^2}{n}.$$

<u>Proof</u>

$$\mathrm{E}(\bar{X})=\mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)=\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(X_i)=\frac{n\mu}{n}=\mu$$

$$\sigma_{\bar{X}}^2=\mathrm{Var}(\bar{X})=\mathrm{Var}(\frac{1}{n}\sum_{i=1}^{n}X_i)=\frac{1}{n^2}\mathrm{Var}(\sum_{i=1}^{n}X_i)\overset{\text{indep.}}{=}\frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i)=\frac{n\sigma^2}{n^2}=\frac{\sigma^2}{n}$$

Note that for the first part of the proof, independence of $X_1,...,X_n$ is not required; however, for the second part independence is essential!
(The independence follows from the statement that it concerns a 'sample from a distribution'.)

Definition 6.6 (*B&E, Eq. 8.2.7*)

Let $X_1,...,X_n$ be a sample. The **sample variance** (Dutch: steekproefvariantie) is defined as

$$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2.$$

Theorem 6.9 (*B&E, Eq.8.2.3*)

Alternative notations for the sample variance S^2 are:

$$S^2=\frac{1}{n-1}\left(\sum_{i=1}^{n}X_i^2-n\bar{X}^2\right) \quad \text{and} \quad S^2=\frac{1}{n-1}\left(\sum_{i=1}^{n}X_i^2-\frac{1}{n}\left(\sum_{i=1}^{n}X_i\right)^2\right)$$

<u>Proof</u>

$$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i^2-2X_i\bar{X}+\bar{X}^2)$$

$$=\frac{1}{n-1}\left(\sum_{i=1}^{n}X_i^2-\sum_{i=1}^{n}2X_i\bar{X}+\sum_{i=1}^{n}\bar{X}^2\right)=\frac{1}{n-1}\left(\sum_{i=1}^{n}X_i^2-2\bar{X}\sum_{i=1}^{n}X_i+n\bar{X}^2\right)$$

$$=\frac{1}{n-1}\left(\sum_{i=1}^{n}X_i^2-2n\bar{X}^2+n\bar{X}^2\right)=\frac{1}{n-1}\left(\sum_{i=1}^{n}X_i^2-n\bar{X}^2\right)$$

The expected value of the sample variance is always equal to the population variance:

Theorem 6.10 (*B&E, Th.8.2.2*)

If $X_1,...,X_n$ is a sample from a distribution with $\mathrm{E}(X)=\mu$ and $\mathrm{Var}(X)=\sigma^2$, then:

$$\mathrm{E}(S^2)=\sigma^2.$$

<u>Proof</u>

Recall first that $\mathrm{E}(X^2)=\sigma^2+\mu^2$ (because $\mathrm{Var}(X)=\sigma^2=\mathrm{E}(X^2)-\mu^2$),
and $\mathrm{E}(\bar{X}^2)=\sigma^2/n+\mu^2$ (because $\mathrm{Var}(\bar{X})=\sigma^2/n=\mathrm{E}(\bar{X}^2)-\mu^2$).

Thus:

$$\mathrm{E}(S^2) = \mathrm{E}\left(\frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right)\right) = \frac{1}{n-1}\left(\mathrm{E}\left(\sum_{i=1}^{n} X_i^2\right) - \mathrm{E}\left(n\,\bar{X}^2\right)\right)$$

$$= \frac{1}{n-1}\left(n\mathrm{E}(X^2) - n\ \mathrm{E}(\bar{X}^2)\right)$$

$$= \frac{1}{n-1}\left(n(\sigma^2+\mu^2) - n\left(\sigma^2\!/_{n} + \mu^2\right)\right) = \frac{n-1}{n-1}\sigma^2 = \sigma^2$$

When a sample is drawn from a Bernoulli distribution with a certain probability p of success, then we can define the following statistic:

Definition 6.7

Let $X_1,...,X_n$ be a sample from a distribution with $X_i \sim \mathrm{BIN}(1, p)$, such that

$$Y = \sum_{i=1}^{n} X_i \sim \mathrm{BIN}(n, p)\ .$$

The **sample proportion** (Dutch: steekproeffractie of steekproefproportie) is defined as

$$\hat{P} = \frac{Y}{n}\ .$$

The sample proportion is in fact just a sample mean. But since it happens so often that we are dealing with a sample from a Bernoulli-distribution, this statistic deserves its own name and treatment.

Remark. In many text books, the sample proportion is notated using a small letter $\hat{p}$ instead of $\hat{P}$. The disadvantage of using a small letter is that there is no longer any distinction in notation between the random variable $\hat{P}$ and the observed sample proportion in a specific sample, for which we will reserve the notation $\hat{p}$.

Theorem 6.11 (*B&E, Ex.8.2.2*)

If $X_1,...,X_n$ is a sample from a distribution with $X_i \sim \mathrm{BIN}(1, p)$, then:

$$\mathrm{E}(\hat{P}) = p \text{ and } \mathrm{Var}(\hat{P}) = \frac{pq}{n}$$

Proof

$$\mathrm{E}(\hat{P}) = \frac{1}{n}\mathrm{E}(\sum_{i=1}^{n} X_i) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(X_i) = \frac{np}{n} = p$$

$$\mathrm{Var}(\hat{P}) = \frac{1}{n^2}\mathrm{Var}(\sum_{i=1}^{n} X_i) \overset{indep.}{=} \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{npq}{n^2} = \frac{pq}{n}$$

6.4.4 Samples from normal distributions

The study of samples from normal distributions deserves special attention. Many results within the field of statistics rely on the assumption of normality of the population distribution. The normal distribution is not only very common, but because of the Central Limit Theorem (CLT, mentioned already in Chapter 4, but only proven in the course Prob. Theory and Statistics 3) it becomes even more important. If the sample size is large enough, the CLT tells us that the sample mean will follow approximately a normal distribution, and because of that also a number of other results will hold approximately, even when sampling from other distributions.

The following result has already been discussed in the previous course, but will be proven here again:

Theorem 6.12 (≈*B&E, Th.8.3.1*)

Let $X_1, \ldots, X_n$ be a sample from a normal distribution, with $X_i \sim \mathrm{N}(\mu, \sigma^2)$, then the sample mean $\bar{X}$ is also normally distributed with expectation μ and variance σ^2/n, so:

$$\bar{X} \sim \mathrm{N}(\mu,\ \sigma^2/n).$$

This can also be written equivalently as: $\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}} = \sqrt{n}\dfrac{\bar{X}-\mu}{\sigma} \sim \mathrm{N}(0,\ 1)$.

<u>*Proof*</u>

$$M_{\bar{X}}(t) = \mathrm{E}[e^{t\bar{X}}] = \mathrm{E}\left[e^{\frac{t}{n}\sum X_i}\right]$$

$$= \mathrm{E}\left[e^{\frac{t}{n}X_1} \cdot e^{\frac{t}{n}X_2} \cdots e^{\frac{t}{n}X_n}\right] \overset{\text{indep.}}{=} \mathrm{E}\left[e^{\frac{t}{n}X_1}\right] \cdot \mathrm{E}\left[e^{\frac{t}{n}X_2}\right] \cdot \ldots \cdot \mathrm{E}\left[e^{\frac{t}{n}X_n}\right]$$

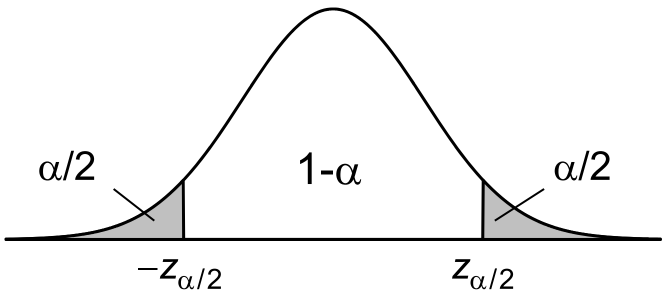
$$= M_{X_1}(t/n) \cdot M_{X_2}(t/n) \cdot \ldots \cdot M_{X_n}(t/n) = \left(\exp\left[\frac{\mu t}{n} + \frac{1}{2}\left(\frac{\sigma t}{n}\right)^2\right]\right)^n$$

$$= \exp\left[\mu t + \frac{1}{2}\frac{(\sigma t)^2}{n}\right]$$

This is indeed the mgf of a normal random variable with expectation μ and variance σ^2/n. The second part of the theorem follows from the process of standardising a normal random variable (see Theorem 4.11 in the previous reader).

This theorem gives (of course) the same expectation and variance for the sample mean as Theorem 6.8 does, which is applicable to all population distributions. But Theorem 6.12 takes this one step further and uses the normality of the population distribution to draw the conclusion that the sample mean has a normal distribution as well.

Since it follows from Theorem 6.12 that the *distribution of* $\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ does not depend on μ and σ, we can, for any arbitrary value of α (between 0 and 1), find a lower limit and an upper limit such that $P(\text{lower limit} < \dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}} < \text{upper limit}) = 1-\alpha$. Although we could choose those limits in many different ways, we usually choose those which are based on the idea of "equal tails" as indicated in the figure below.



Thus

$$P(-z_{\alpha/2} < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < z_{\alpha/2}) = 1-\alpha$$

from which it follows that

$$P(\mu - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \overline{X} < \mu + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

So we know that with a probability of 1– α, the sample mean will fall within the interval

$$(\mu - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \mu + z_{\alpha/2}\frac{\sigma}{\sqrt{n}})$$

which is sometimes referred to as the 100(1–α)%-*prediction interval* for $\overline{X}$. Note that this interval can only be determined in case both μ and σ have known values.

Remark. We can also consider a 100(1–α)%-prediction interval for $\dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}}$. Because this random variable has a standard normal distribution, the prediction interval is then simply $(-z_{\alpha/2},\ z_{\alpha/2})$.

Example 6.18

Assume that the heights of adult Dutch males is normally distributed with a mean of 182 cm and a standard deviation of 7 cm. We will determine here a 90%-prediction interval for the sample mean with a sample size of 12. We know that:

$$P(182 - \frac{7}{\sqrt{12}}1.645 < \overline{X} < 182 + \frac{7}{\sqrt{12}}1.645) = 0.90\ .$$

So the 90%-prediction interval will be

$$(182 - \frac{7}{\sqrt{12}}1.645,\ 182 + \frac{7}{\sqrt{12}}1.645) = (178.676,\ 185.324)$$

Or: with a probability of 90%, the sample mean will lie between 178.676 cm and 185.324 cm. ◄

The following two theorems are essential in order to be able to prove in Chapter 7 the very important Theorem 7.1 and Theorem 7.2.

Theorem 6.13 (*B&E, Eq.8.3.13*)

Let $X_1, \ldots, X_n$ be a sample from a normal distribution, with $X_i \sim N(\mu, \sigma^2)$. Then:

$$\sum_{i=1}^{n}\frac{(X_i-\mu)^2}{\sigma^2} \sim \chi^2(n)$$

and $n\dfrac{(\overline{X}-\mu)^2}{\sigma^2} \sim \chi^2(1)$

Proof

From Theorem 4.11 (standardising), we know that $\dfrac{X_i-\mu}{\sigma} \sim N(0, 1)$.

Using Theorem 4.14, which states that the square of a standard normal random variable has a χ^2-distribution with 1 degree of freedom, we obtain $\dfrac{(X_i-\mu)^2}{\sigma^2} \sim \chi^2(1)$.

We have also seen before (see for example section 6.2.2) that the sum of independent chi-square distributed random variables is again chi-square distributed, where the number of degrees of freedom is summed, which proves the first part.

For the second part: it follows from Theorem 6.12 that $\dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$. Using Theorem 4.14, we get the result that $\dfrac{(\overline{X}-\mu)^2}{\sigma^2/n} \sim \chi^2(1)$.

Theorem 6.14 (*B&E, Th.8.3.6*)

If $X_1, \ldots, X_n$ is a sample from a normal distribution, with $X_i \sim N(\mu, \sigma^2)$, then

1. $\bar{X}$ and S^2 are independent random variables
2. $\dfrac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}{\sigma^2} = \dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$

<u>*Proof*</u>

1. The joint pdf of $X_1,\ldots,X_n$ is (with $-\infty < x_i < \infty$, for $i = 1,\ldots,n$)

$$f_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = \prod_{i=1}^{n} f(x_i) = \frac{1}{(2\pi)^{n/2}\sigma^n}\exp\left[-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2\right]$$

$$= \frac{1}{(2\pi)^{n/2}\sigma^n}\exp\left[-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(x_i-\bar{x})^2 + n(\bar{x}-\mu)^2\right)\right]$$

$x_i^2 - 2x_i\bar{x} + \bar{x}^2$

$\sum_{i=1}^{n}(x_i-\bar{x})^2$

Now define the following transformation:

$y_1 = \bar{x},\ y_i = x_i - \bar{x}$ for $i = 2,\ldots,n$ (such that $-\infty < y_i < \infty$, for $i = 1,\ldots,n$)

$\sum_{i=1}^{n} y_i^2$

Because $\sum_{i=1}^{n}(x_i - \bar{x})$ is by definition of the sample mean equal to 0, it follows that

$= y_1^2 + \sum_{i=2}^{n} y_i^2$

$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

$$x_1 - \bar{x} = -\sum_{i=2}^{n}(x_i - \bar{x}) = -\sum_{i=2}^{n} y_i$$

$\hookrightarrow (x_1 - \bar{x})^2$

$$\Rightarrow \quad x_1 = y_1 - \sum_{i=2}^{n} y_i \quad \text{and} \quad x_i = y_i + y_1 \quad (\text{for } i = 2,\ldots,n)$$

$\sum_{i=1}^{n}(x_i-\bar{x})^2$

$x_1 = y_i + \bar{x}$

This results in the Jacobian:

$\frac{\partial x_1}{\partial y_1}$ - - -

$\sum_{i=1}^{n} x_i^2 + n\bar{x}^2$

$$J = \begin{vmatrix} 1 & -1 & -1 & \cdots & -1 \\ 1 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ 1 & 0 & 0 & \cdots & 1 \end{vmatrix} = \begin{vmatrix} n & 0 & 0 & \cdots & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ 1 & 0 & 0 & \cdots & 1 \end{vmatrix} = n$$

$\frac{\partial x_k}{\partial y_k}$ - - -

y_1^2

$$\Rightarrow \quad f_{Y_1,\ldots,Y_n}(y_1,\ldots,y_n) = \frac{n}{(2\pi)^{n/2}\sigma^n}\exp\left[-\frac{1}{2\sigma^2}\left(\left(-\sum_{i=2}^{n} y_i\right)^2 + \sum_{i=2}^{n} y_i^2 + n(y_1-\mu)^2\right)\right]$$

Looking at this pdf, it becomes clear that it can be factorised into the product of a function of $y_1(=\bar{x})$ and a function of all other $y_i(=x_i - \bar{x})$ for $i = 2,\ldots,n$. Application of Theorem 5.10 shows that this means that $\bar{X}$ is independent from $(X_2 - \bar{X},\ldots,X_n - \bar{X})$. And since

$$(n-1)S^2 = \sum_{i=1}^{n}(X_i-\bar{X})^2 = (X_1-\bar{X})^2 + \sum_{i=2}^{n}(X_i-\bar{X})^2 = \left(-\sum_{i=2}^{n}(X_i-\bar{X})\right)^2 + \sum_{i=2}^{n}(X_i-\bar{X})^2$$

we can see that S^2 is a function of $(X_2 - \bar{X},\ldots,X_n - \bar{X})$, which proves that $\bar{X}$ is independent of S^2.

2. Note that we can write:

$$\sum_{i=1}^{n}\left(x_i-\mu\right)^2 = \sum_{i=1}^{n}(x_i - \bar{x} + \bar{x} - \mu)^2$$

$$= \sum_{i=1}^{n}(x_i-\bar{x})^2 + 2(\bar{x}-\mu)\sum_{i=1}^{n}(x_i-\bar{x}) + \sum_{i=1}^{n}(\bar{x}-\mu)^2$$

$$= \sum_{i=1}^{n} (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2 \quad \text{(because } \sum_{i=1}^{n} (x_i - \bar{x}) = 0 \text{ !)}$$

Therefore:

$$\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{\sigma^2} + \frac{n(\bar{x} - \mu)^2}{\sigma^2} = \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{\sigma^2},$$

$$\Rightarrow \quad \frac{(n-1)S^2}{\sigma^2} + \frac{n(\bar{X} - \mu)^2}{\sigma^2} = \sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma^2}$$

$\bar{X}$ & S^2.

The left-hand side of the last equation is the sum of two independent random variables (see first part of this theorem), which means that we can apply Theorem 5.11 (1c):

$$\Rightarrow \quad M\left[\frac{(n-1)S^2}{\sigma^2}; t\right] \cdot M\left[\frac{n(\bar{X} - \mu)^2}{\sigma^2}; t\right] = M\left[\sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma^2}; t\right]$$

Theorem 6.13 tells us that the right-hand side is equal to $(1-2t)^{-n/2}$, and that the second factor on the left-hand side is equal to $(1-2t)^{-1/2}$. Thus we get:

$$M\left[\frac{(n-1)S^2}{\sigma^2}; t\right] \cdot (1-2t)^{-1/2} = (1-2t)^{-n/2}$$

$$\Rightarrow \quad M\left[\frac{(n-1)S^2}{\sigma^2}; t\right] = (1-2t)^{-(n-1)/2}.$$

This shows that $\dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$.

The result of the second part of this theorem can be used to derive a prediction interval for the sample variance:

$$P\left(\chi^2_{n-1;\,1-\alpha/2} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{n-1;\,\alpha/2} \right) = 1 - \alpha$$

$$\Rightarrow \quad P\left(\frac{\sigma^2}{n-1}\chi^2_{n-1;\,1-\alpha/2} < S^2 < \frac{\sigma^2}{n-1}\chi^2_{n-1;\,\alpha/2} \right) = 1 - \alpha$$

Thus $\left(\dfrac{\sigma^2}{n-1}\chi^2_{n-1;\,1-\alpha/2}, \dfrac{\sigma^2}{n-1}\chi^2_{n-1;\,\alpha/2} \right)$ is a 100(1–α)%-prediction interval for S^2.

We mention here as well already Theorem 7.1 and Theorem 7.2 (in order to show all important result for samples from normal distributions in one place).

If $X_1, X_2, \ldots, X_n$ are independent random variables, with $X_i \sim \mathrm{N}(\mu, \sigma^2)$, then:

$$T_{n-1} = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

If $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$ are independent random variables, with $X_i \sim \mathrm{N}(\mu_1, \sigma_1^2)$ and $Y_i \sim \mathrm{N}(\mu_2, \sigma_2^2)$, then:

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{S_1^2\sigma_2^2}{S_2^2\sigma_1^2} \sim F(n_1 - 1, n_2 - 1)$$

6.5 Order Statistics

(B&E, pages 214-221)

6.5.1 Distribution of the minimum or maximum of n *random variables*

Often, the distribution of either the minimum or the maximum of *n* random variables needs to be determined. We can apply the techniques which have been discussed earlier in the chapter.

Example 6.19

An instrument contains two components, with lifespans which are independently and exponentially distributed with expected values of 1 and 0.5 year respectively, so $X_1 \sim \text{EXP}(1)$ and $X_2 \sim \text{EXP}(2)$. Assume that the instrument will continue to function as long as at least one of the two components functions. We would like to know the distribution of the total time *V* the instrument will function. It is clear that $V = \max(X_1, X_2)$ and that the support of *V* is the positive real line. Using the CDF-method, we can derive:

$$F_V(v) = P(V \le v) = P(\max(X_1, X_2) \le v)$$
$$= P(X_1 \le v,\ X_2 \le v) = \int_0^v\int_0^v f_{X_1,X_2}(x_1,x_2)\,dx_2\,dx_1$$

(Note the important third step: the maximum of X_1 and X_2 is less than or equal to *v* if and only if both $X_1 \le v$ and $X_2 \le v$).

Instead of evaluating the above integral, we can use the information that X_1 and X_2 are independent:

$$F_V(v) = P(X_1 \le v,\ X_2 \le v) \overset{\text{indep.}}{=} P(X_1 \le v)\cdot P(X_2 \le v) = F_{X_1}(v)\cdot F_{X_2}(v)$$

From $X_1 \sim \text{EXP}(1)$ and $X_2 \sim \text{EXP}(2)$, it follows that $F_{X_1}(v) = 1 - e^{-1v}$ and $F_{X_2}(v) = 1 - e^{-2v}$

Then we get:

$$F_V(v) = (1 - e^{-1v})\cdot(1 - e^{-2v}) = 1 - e^{-v} - e^{-2v} + e^{-3v}$$

which after differentiation results in the pdf

$$f_V(v) = e^{-v} + 2e^{-2v} - 3e^{-3v} \qquad \text{for } v > 0$$

This is not one of the known distributions (it would be a mistake to think that this represents the distribution of $Y_1 + Y_2 - Y_3$, where $Y_1 \sim \text{EXP}(1)$, $Y_2 \sim \text{EXP}(2)$ and $Y_3 \sim \text{EXP}(3)$!). But we can see very simply that the expected value of *V* is equal to 1 + 0.5 – 0.333 = 1.167, or 14 months (check!). ◄

Example 6.20

Consider again an instrument with two components, with the same lifespans as in the previous example. However, now assume that the instrument only works properly as long as both components are still functioning. The lifespan *W* of the instrument is then: $W = \min(X_1, X_2)$. Note that it is *not* true that $\min(X_1, X_2) \le w$ if and only if both $X_1 \le w$ and $X_2 \le w$. But we can say that $\min(X_1, X_2) > w$ if and only if both $X_1 > w$ and $X_2 > w$. Thus:

$$F_W(w) = P(W \le w) = P(\min(X_1, X_2) \le w) = 1 - P(\min(X_1, X_2) > w)$$
$$= 1 - P(X_1 > w,\ X_2 > w) \overset{\text{indep.}}{=} 1 - P(X_1 > w)\cdot P(X_2 > w)$$
$$= 1 - (1 - F_{X_1}(w))\cdot(1 - F_{X_2}(w))$$
$$= 1 - e^{-1w}\cdot e^{-2w} = 1 - e^{-3w}$$

Which results in the pdf of *W*:

$$f_W(w) = 3e^{-3w} \qquad \text{for } w > 0$$

In other words, the lifespan of the instrument follows an exponential distribution with expected value of 1/3 year. ◄

For the remainder of the section, we will limit ourselves to cases where the random variables X_1 to X_n form a sample from a distribution with pdf $f(x)$, so $f_{X_1,\dots,X_n}(x_1,\dots,x_n) = f(x_1)\cdot f(x_2)\cdot\ldots\cdot f(x_n)$.

Theorem 6.15 (*B&E, Eq.6.5.7/8, ±Th.6.5.2*)

Let $X_1,\dots, X_n$ be a sample from a distribution with pdf $f(x)$ and CDF $F(x)$.

The CDF of $V = \max(X_1,\dots, X_n)$ is $F_V(v) = [F(v)]^n$, and the CDF of $W = \min(X_1,\dots, X_n)$ is $F_W(w) = 1-[1-F(w)]^n$.

If the distribution of X is *continuous*, then the pdf of V is $f_V(v) = n[F(v)]^{n-1} f(v)$, and the pdf of W is $f_W(w) = n[1-F(w)]^{n-1} f(w)$.

Proof

$$
\begin{aligned}
F_V(v) &= P(V \le v) = P(\max(X_1,\dots,X_n) \le v)\\
&= P(X_1 \le v,\ X_2 \le v,\dots,X_n \le v)\\
&= P(X_1 \le v)\cdot P(X_2 \le v)\cdot\ldots\cdot P(X_n \le v)\\
&= \left[P(X \le v)\right]^n = \left[F(v)\right]^n
\end{aligned}
$$

$$
\begin{aligned}
F_W(w) &= P(W \le w) = P(\min(X_1,\dots,X_n) \le w)\\
&= 1 - P(\min(X_1,\dots,X_n) > w)\\
&= 1 - P(X_1 > w)\cdot P(X_2 > w)\cdot\ldots\cdot P(X_n > w)\\
&= 1 - \left[P(X > w)\right]^n = 1 - \left[1 - F(w)\right]^n
\end{aligned}
$$

If the sampling was done from a continuous distribution, we can then get:

$$f_V(v) = \frac{dF_V(v)}{dv} = n[F(v)]^{n-1} f(v)$$

and

$$f_W(w) = \frac{dF_W(w)}{dw} = -n[1-F(w)]^{n-1}(0 - f(w)) = n[1-F(w)]^{n-1} f(w)$$

Example 6.21

Throw three fair dice. We will determine the CDF's of the highest and of the lowest number of dots on the three dice. First, we note that the CDF of the number of dots when throwing one single die is $F(x) = \lfloor x \rfloor / 6$, where $\lfloor x \rfloor$ denotes the Entier-function, which is the largest integer less than or equal to x (in many programming languages called the 'Floor of x'). So, for example $F(2) = \lfloor 2 \rfloor / 6 = 2/6$ and $F(2.9) = \lfloor 2.9 \rfloor / 6 = 2/6$, which is indeed correct.

We write X_1, X_2 and X_3 for the number of dots on the three dice. From Theorem 6.15, it follows that the CDF of the maximum $V = \max(X_1, X_2, X_3)$ is

$F_V(v) = (\lfloor v \rfloor / 6)^3$ (for $1 \le v \le 6$).

The CDF of the minimum $W = \min(X_1, X_2, X_3)$ is: $F_W(w) = 1-(1-\lfloor w \rfloor / 6)^3$ (for $1 \le w \le 6$). Since V and W are discrete random variables, we can find the pdf only by determining where the CDF jumps upwards. For example:

$$P(\max(X_1, X_2, X_3) = 5) = P(V = 5) = P(V \le 5) - P(V \le 4) = F_V(5) - F_V(4)$$
$$= (5/6)^3 - (4/6)^3 = 61/216$$ ◄

Example 6.22

Each of three light bulbs has an exponentially distributed lifespan (X_1, X_2 and X_3 respectively, mutually independent) with an expectation of 24 months. We want to find the pdf of the time until one of those three bulbs stops functioning, which is the pdf of $Y = \min(X_1, X_2, X_3)$. Then:

$$f_Y(y) = 3(1 - F_X(y))^2 \cdot f_X(y) = 3(1-(1-e^{-\frac{1}{24}y}))^2 \cdot \tfrac{1}{24} e^{-\frac{1}{24}y} = \tfrac{1}{8} e^{-\frac{1}{8}y}$$

Thus Y is exponentially distributed with an expected value of 8 months. ◄

6.5.2 *Distribution of all order statistics*

In the previous section, we paid attention to the distributions of the minimum and of the maximum. But sometimes we are interested in the distributions of, for example, the second or third observation in size. In this section, we will discuss samples taken from *continuous* distributions only. After ordering a sample $X_1, \ldots, X_n$ in increasing magnitude, we obtain the ordered random variables Y_1 to Y_n, such that:

$$Y_1 = \min(X_1, \ldots, X_n) < Y_2 < \cdots < Y_n = \max(X_1, \ldots, X_n)$$

The notation $X_{i:n}$ is often used instead of Y_i. Say for example that our observed sample is 3, -1, 0, -2, 17 and 8, then we get $y_1 (= x_{1:n}) = -2,\ y_2 (= x_{2:n}) = -1,\ y_3 = 0,\ y_4 = 3,\ y_5 = 8,\ y_6 = 17$.

The joint pdf of $Y_1, \ldots, Y_n$ is not the same as the joint pdf of $X_1, \ldots, X_n$. When, as in the example above, we have 6 observations, then there will be 6! different permutations of $x_1, \ldots, x_6$ (all equally likely, since they form a sample from a distribution) which will result in the same values for $y_1, \ldots, y_6$. This is reflected in the following theorem. (Note that $Y_1, \ldots, Y_n$ are not mutually independent anymore!)

Theorem 6.16 (*B&E, Th.6.5.1*)

If $X_1, \ldots, X_n$ is a sample from a continuous distribution with pdf $f(x)$, and $Y_1, \ldots, Y_n$ is the ordered sample, then:

$$f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n) = n!\, f(y_1) \cdot f(y_2) \cdot \cdots \cdot f(y_n) \qquad \text{for} \quad y_1 < y_2 < \cdots < y_n .$$

Example 6.23

Let X_1, X_2 and X_3 be a sample from the uniform distribution, with $X_i \sim$ UNIF(0,1).

Then it follows that: $f_{Y_1,Y_2,Y_3}(y_1, y_2, y_3) = 3!(1) \cdot (1) \cdot (1) = 6 \qquad \text{for } 0 < y_1 < y_2 < y_3 < 1$.

Note this is the joint pdf we already encountered before in Example 5.15 and Example 5.42. ◄

Example 6.24

Let X_1, X_2 and X_3 be a sample from a distribution, with

$$f_X(x) = 2x \qquad (\text{for } 0 < x < 1).$$

Thus: $f_{Y_1,Y_2,Y_3}(y_1, y_2, y_3) = 3!(2y_1) \cdot (2y_2) \cdot (2y_3) \qquad \text{for } 0 < y_1 < y_2 < y_3 < 1$.

Of course, we can use this joint pdf to find any of the marginal pdf's . For example:

$$\begin{aligned} f_{Y_2}(y) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{Y_1,Y_2,Y_3}(y_1, y, y_3)\, dy_1 dy_3 \\ &= \int_y^1 \int_0^y 3!(2y_1) \cdot (2y) \cdot (2y_3)\, dy_1 dy_3 \\ &= 3!(2y) \int_y^1 2y_3 \int_0^y 2y_1\, dy_1 dy_3 = 3!(2y) y^2 \int_y^1 2y_3 dy_3 \end{aligned}$$

$$= 3!(2y)y^2(1-y^2) = 12(y^3 - y^5) \qquad \text{for } 0 < y < 1$$ ◄

We can also find a general expression for the CDF of any of the order statistics:

Theorem 6.17 (*B&E, Th.6.5.3*)

For any sample from a distribution (both continuous or discrete) with CDF $F(x)$, the CDF of the *k*-th order statistic is given by:

$$F_{Y_k}(y) = \sum_{j=k}^{n} \binom{n}{j} [F(y)]^j \, [1-F(y)]^{n-j}$$

Proof

The main idea in this proof is recognising that we can use the binomial distribution.

$$F_{Y_k}(y) = P(Y_k \le y) = P(\text{at least } k \text{ of the } X_i\text{'s are } \le y)$$

$$= \sum_{j=k}^{n} P(\text{exactly } j \text{ of the } X_i\text{'s are } \le y)$$

$$= \sum_{j=k}^{n} \binom{n}{j} [F(y)]^j \, [1-F(y)]^{n-j}$$

(because the number of observations less than or equal to *y* will have a binomial distribution with parameter *n* and a probability of success equal to $P(X_i \le y) = F(y)$).

In case the sample is taken from a *continuous* distribution, we can now find the pdf of Y_k :

Theorem 6.18 (*B&E, Th.6.5.2*)

For any sample from a continuous distribution with CDF $F(x)$, the pdf of the *k*-th order statistic Y_k is given by:

$$f_{Y_k}(y) = \frac{n!}{(k-1)!(n-k)!} [F(y)]^{k-1} \, [1-F(y)]^{n-k} \, f(y)$$

Proof

$$f_{Y_k}(y) = \frac{d}{dy} F_{Y_k}(y)$$

$$= \sum_{j=k}^{n} \binom{n}{j} j \, f(y)[F(y)]^{j-1} \, [1-F(y)]^{n-j} - \sum_{j=k}^{n} \binom{n}{j} (n-j) \, f(y)[F(y)]^j \, [1-F(y)]^{n-j-1}$$

$$= nf(y) \sum_{j=k}^{n} \binom{n-1}{j-1} [F(y)]^{j-1} \, [1-F(y)]^{n-j} - nf(y) \sum_{j=k}^{n} \binom{n-1}{j} [F(y)]^j \, [1-F(y)]^{n-j-1}$$

$$= nf(y) \sum_{l=k-1}^{n-1} \binom{n-1}{l} [F(y)]^l \, [1-F(y)]^{n-l-1} - nf(y) \sum_{j=k}^{n} \binom{n-1}{j} [F(y)]^j \, [1-F(y)]^{n-j-1}$$

Now almost all terms cancel out, with only two remaining:

$$= nf(y) \binom{n-1}{k-1} [F(y)]^{k-1} \, [1-F(y)]^{n-k} - nf(y) \binom{n-1}{n} [F(y)]^n \, [1-F(y)]^{n-n-1}$$

$$= f(y) \frac{n!}{(k-1)!(n-k)!} [F(y)]^{k-1} \, [1-F(y)]^{n-k} \qquad \left(\text{because } \binom{n-1}{n} = 0\right)$$

Example 6.25

Let X_1, X_2 and X_3 be a sample from the uniform distribution, with $X_i \sim$ UNIF(0,1). We will use the previous theorem to find the pdf of the midmost observation, so $f_{Y_2}(y)$. Note first that in this case we have $F(x) = x$. Then we get:

$$f_{Y_2}(y) = \frac{3!}{(2-1)!(3-2)!}[F(y)]^{2-1}[1-F(y)]^{3-2} f(y)$$
$$= 6y[1-y]\cdot 1$$
$$= 6y - 6y^2 \quad \text{for } 0 < y < 1$$

◄

In a similar way, we can also determine the joint pdf of two arbitrary random variables Y_i and Y_j from the ordered sample.

Theorem 6.19 *(B&E, Eq.6.5.4)*

For any sample from a continuous distribution with CDF $F(x)$, and with $j > i$:

$$f_{Y_i,Y_j}(y_i, y_j) = \frac{n!}{(i-1)!(j-i-1)!(n-j)!} f(y_i) f(y_j)[F(y_i)]^{i-1}[F(y_j)-F(y_i)]^{j-i-1}[1-F(y_j)]^{n-j}$$

(for $y_i < y_j$)

Proof

(We only give here a rough sketch of the proof; it may be skipped):
First, we focus on the probability that out of the n observations, exactly 1 falls within the interval $(y_i, y_i + dy_i)$, and another in $(y_j, y_j + dy_j)$, and $i - 1$ are less than y_i, and $j - i - 1$ are between $y_i + dy_i$ and y_j and the remaining $n - j$ observations are greater than $y_j + dy_j$. This means we are dealing with a multinomial distribution, and the requested probability is

$$\frac{n!}{1!\cdot 1!\cdot (i-1)!\cdot (j-i-1)!\cdot (n-j)!} \int_{y_i}^{y_i+dy_i} f(s)ds \int_{y_j}^{y_j+dy_j} f(t)dt\,[F(y_i)]^{i-1}$$
$$\times [F(y_j) - F(y_i + dy_i)]^{j-i-1}[1 - F(y_j + dy_j)]^{n-j}$$

Logically, this probability should be equal to:

$$\int_{y_i}^{y_i+dy_i} \int_{y_j}^{y_j+dy_j} f_{Y_i,Y_j}(u,v)dudv$$

When we take the limit of both dy_i and dy_j to 0, then we see that

$$\int_{y_i}^{y_i+dy_i} \int_{y_j}^{y_j+dy_j} f_{Y_i,Y_j}(u,v)dudv \approx f_{Y_i,Y_j}(y_i, y_j)dy_i dy_j$$

Similarly, we know that $\int_{y_i}^{y_i+dy_i} f_X(s)ds \approx f_X(y_i)dy_i$ and $\int_{y_j}^{y_j+dy_j} f_X(s)ds \approx f_X(y_j)dy_j$. When we substitute this into the derived equality, we arrive at the desired result.

Using both the previous theorems and the transformation method from section 6.1.2 , we can determine now as well the distributions of other statistics like the **sample range** $R = Y_n - Y_1$ (Dutch:

steekproefbereik), and the **sample median**, which is defined as $Y_{(n+1)/2}$ if n is odd, and $(Y_{n/2}+Y_{n/2+1})/2$ is n is even).

Example 6.26

Consider $X_1,...,X_n$ to be a sample from a distribution with $f(x)=2x$ (for $0<x<1$), such that $F(x)=x^2$ (for $0<x<1$).

We will now determine the pdf of the sample range. First, we can use Theorem 6.19 to write the joint pdf of Y_1 ($=\min(X_1,...,X_n)$) and Y_n ($=\max(X_1,...,X_n)$):

$$f_{Y_1,Y_n}(y_i,y_n)=\frac{n!}{(n-2)!}(2y_1)(2y_n)\,[y_n^2-y_1^2]^{n-2} \qquad \text{for } 0<y_1<y_n<1.$$

Next, we apply the transformation method with $R=Y_n-Y_1$, $S=Y_1$. The inverse transformation is $y_1=s$, $y_n=r+s$ and $|J|=1$, such that the joint pdf of *R* and *S* becomes:

$$f_{R,S}(r,s)=\frac{4n!}{(n-2)!}s(r+s)\,[r^2+2rs]^{n-2} \qquad \text{for } 0<s<1-r,\ 0<r<1$$

In the last step, we determine the marginal density of *R*:

$$f_R(r)=\int_{-\infty}^{\infty} f_{R,S}(r,s)ds=\int_0^{1-r}\frac{4n!}{(n-2)!}s(r+s)\,[r^2+2rs]^{n-2}ds \qquad \text{for } 0<r<1$$

If, for example, we take $n = 2$, then:

$$f_R(r)=\int_0^{1-r} 8s(r+s)\,ds=\frac{4}{3}(r+2)(1-r^2) \qquad \text{for } 0<r<1$$ ◄

6.6 Problems

6.1 (B&E 6.1) Given is the random variable *X* with pdf $f_X(x) = 4x^3$ (for $0 < x < 1$). Use the CDF method to determine the pdf of each of the following random variables:

a $Y=X^4$

b $W=e^X$

c $Z=\ln X$

d $U=(X-0.5)^2$

6.2 (B&E 6.8) Rework the previous exercise using transformation methods.

6.3 Let *A* be the area defined by $0 \le x \le 1$ and $0 \le y \le x^2$, and let the joint density be given by $f_{X,Y}(x, y) = kxy$ for $(x, y) \in A$, and 0 otherwise.

a Draw the support of (*X*, *Y*) and calculate *k*.

b Are *X* and *Y* independent?

c Calculate the marginal probability distributions of *X* and *Y*.

d Determine using the CDF method the pdf of $Z = XY$.

6.4 For the random variables *X* and *Y* the simultaneous density is equal to ½ for $x, y \ge 0$ and $x + 4y \le 4$ (outside the triangle the density equals 0). Determine using the CDF method the CDF and the pdf of the variable $W = Y / X$.

6.5 The variables *X* and *Y* have a simultaneous distribution defined by

$$f_{X,Y}(x,y)=\frac{2x}{25} \quad \text{on the area: } 0<x<5 \text{ and } 0<y<3-0.6x$$

a Are *X* and *Y* independent?

b Determine using the CDF method the pdf of $W = Y / X$.

6.6 The joint pdf of *X* and *Y* is $f_{X,Y}(x, y) = 4xy$ for $0 \le x < 1$ and $0 \le y < 1$.
Define $Z = XY$. Determine using the CDF method the CDF and the pdf of *Z*.

6.7 (Problem 5.17 continued). The variables X and Y have a simultaneous distribution defined by

$f_{X,Y}(x, y) = 3y/2$ on the area: $0 < x < 4$ and $0 < y < 1 - \frac{1}{4}x$

Determine the pdf of $W = X + Y$ and calculate the expectation of $X + Y$ and check the result with the expected values found in Problem 5.17.

6.8 The pdf's of X and Y are:

$f_X(x) = 2x$ for $0 \le x < 1$

$f_Y(y) = 2(1 - y)$ for $0 \le y < 1$

Furthermore, X and Y are independent. Define $Z = X + Y$. Determine using the CDF method the CDF and the pdf of Z.

6.9 The independent random variables X and Y are both UNIF(0, 1) distributed. The random variable U is defined as $U = Y \cdot X^2$. Determine the conditional pdf of U given $Y = y$.

6.10 The joint pdf of X and Y is given by:

$f_{X,Y}(x, y) = \theta^2 e^{-\theta(x+y)}$ for $0 < x, 0 < y$.

a Use the transformation method to determine the simultaneous pdf of $V = X/Y$ and $W = X + Y$.
b Determine the marginal pdf of V as well as of W.
c Are V and W independent?

6.11 (B&E 6.14) The simultaneous pdf of X and Y is given by:

$f_{X,Y}(x, y) = 4e^{-2(x+y)}$ for $0 < x, 0 < y$.

a Use the transformation method to determine the distribution of $W = X + Y$.
b Use the transformation method to determine the simultaneous pdf of $U = X/Y$ and $V = X$.
c Find the marginal pdf of U.

6.12 (B&E 6.16) X_1 and X_2 denote a random sample of size two from a distribution with pdf $f(x) = 1/x^2$ for $1 < x$.

a Use the transformation method to determine the simultaneous pdf of $U = X_1 X_2$ and $V = X_1$.
b Find the marginal pdf of U.

6.13 The three random variables X, Y and Z have the following simultaneous pdf:

$f_{X,Y,Z}(x, y, z) = x^2 e^{-x(1+y+z)}$ for $x, y, z > 0$

Determine the simultaneous pdf of U, V and W where $U = X$, $V = XY$ and $W = XZ$. Are U, V and W mutually independent?

6.14 (B&E 6.17) X_1 and X_2 denote a random sample of size 2 from a Gamma-distribution, $X_i \sim$ GAM(2, 1/2).

a Find the pdf of $Y = \sqrt{X_1 + X_2}$.
b Find the pdf of $W = X_1 / X_2$ using the transformation method.

6.15 (B&E 6.18) The simultaneous pdf of X and Y is given by: $f_{X,Y}(x, y) = e^{-y}$ for $0 < x < y$.

a Use the transformation method to determine the simultaneous pdf of $S = X + Y$ and $T = X$.
b Find the marginal pdf of T.
c Find the marginal pdf of S.

6.16 (B&E 6.21) The simultaneous pdf of X and Y is given by:

$f_{X,Y}(x, y) = 2(x + y)$ for $0 < x < y < 1$.

a Use the transformation method to determine the simultaneous pdf of $S = X$ and $T = XY$.
b Find the marginal pdf of T.

6.17 (B&E 8.5) A new electronical component is placed in service and nine spare parts are available. The times to failure in days are independent exponential variables, $T_i \sim$ EXP(1/100).

a What is the distribution of $\sum_{i=1}^{10} T_i$?
b What is the probability that successful operation can be maintained for at least 1.5 years? Hint: use Theorem 4.15 (Prob. Th & St 1).
c How many spare parts would be needed to be 95% sure of successful operation for at least two years?

6.18 (B&E 6.35) Let X_1 and X_2 denote a sample from the exponential distribution, $X_i \sim \text{EXP}(\lambda)$, and let $Y = X_1 - X_2$.

a Determine the moment generating function of Y.

b What is the distribution of Y?

6.19 Let $X_1, \ldots, X_4$ denote a sample from a standard normal distribution, $X_i \sim \text{N}(0,\ 1)$.

a Determine the moment generating function of $V = X_1^2 + X_2^2$. Which exponential distribution has V?

b Determine the pdf of $U = X_1^2 + X_2^2 - X_3^2 - X_4^2$. (Hint: use mgf's, and the list of continuous distributions at the beginning of the book of Bain and Engelhardt).

6.20 (B&E 5.25) Given are the independent r.v.'s $X_1 \sim \text{N}(\mu_1,\ \sigma_1^2)$ and $X_2 \sim \text{N}(\mu_2,\ \sigma_2^2)$. Define $Y_1 = X_1$ and $Y_2 = X_1 + X_2$.

a Show that Y_1 and Y_2 are bivariate normal distributed. (Hint: use moment generating functions).

b What are the expectations, the variances and the coefficient of correlation of Y_1 and Y_2?

c Find the conditional distribution of Y_2 given $Y_1 = y_1$.

6.21 Let $X_1, X_2, \ldots$ denote a series of independent and identical distributed random variables, each with mgf $M_X(t)$. Define the sum $S = \sum_{i=1}^{N} X_i$, where N denotes a discrete random variable with mgf $M_N(t)$ (and independent of $X_1, X_2, \ldots$). Show that $M_S(t) = M_N(\ln M_X(t))$. (Hint: condition on N, and use Theorem 5.21).

6.22 Assume $S \mid N = n \sim \text{Bin}(n,\ p)$ and $N \sim \text{Poi}(\lambda)$. Use the result of the previous exercise to determine the distribution of S.

6.23 (B&E 8.7) Five independent tasks are to be performed sequentially, where the time (in hours) to complete i-th task is Gamma distributed with parameters 100 and $\kappa_i = 3 + i/3$. What is the probability that it will take less time than 2600 hours to complete all five tasks?

6.24 Assume that the weights (X_1 and X_2) of two bags of potatoes are normal distributed with $\mu_1 = \mu_2 = 5000$ and $\sigma_1 = 28$ and $\sigma_2 = 45$, and independent of each other. Calculate the probability that the first bag has at least 18 gram more than the second bag.

6.25 Assume that the weights (X_1 and X_2) of two bags of potatoes are normal distributed with $\mu_1 = 5078$ and $\mu_2 = 5062$, respectively, and $\sigma_1 = 33$ and $\sigma_2 = 56$ (independent of each other). Calculate the probability that the second bag has at least 10 gram more than the first bag.

6.26 (B&E 8.8) Suppose that $X \sim \chi^2(m)$, $Y \sim \chi^2(n)$ and X and Y are independent. Is $Y - X$ also chi-square distributed if $n > m$?

6.27 (B&E 8.9) Suppose that $X \sim \chi^2(m)$, $S = X + Y \sim \chi^2(m+n)$ and X and Y are independent. Show that $S - X \sim \chi^2(n)$.

6.28 (B&E 8.10) A random sample of size $n = 15$ is drawn from an exponential distribution with expected value θ. Find c so that $P(c\overline{X} < \theta) = 0.95$, where $\overline{X}$ is the sample mean. Hint: Determine first the distribution of $\sum_{i=1}^{15} X_i$, and use Theorem 4.15 (Prob. Th&St 1, see also Example 6.7).

6.29 (B&E 8.12) The distance in feet by which a parachutist misses a target is $D = \sqrt{X_1^2 + X_2^2}$, where X_1 and X_2 are independent with $X_i \sim \text{N}(0,\ 25)$. Find $P(D \leq 12.25)$.

6.30 (B&E 8.14) If $T \sim t(\nu)$, give the distribution of T^2.

6.31 Show that the Cauchy-distribution (see Prob. Th.&St 1) is a special case of the t-distribution.

6.32 If X and Y are independent, with $X \sim \text{N}(0,\ 1)$ and $Y \sim \chi^2(6)$, determine an interval in which $P(X < \sqrt{Y})$ will lie. Hint: Rewrite the required probability into a probability regarding a t-distributed variable.

6.33 If X and Y are independent, with $X \sim \chi^2(15)$ and $Y \sim \chi^2(30)$, find $P(X < Y)$ by using the F-distribution.

6.34 Consider a series of observations $y_1, y_2, y_3, \ldots, y_n$, ordered by increasing size, which were taken from a distribution with a certain CDF F. In a graph two series are plotted against each other:

$$y_i \text{ and } F^{-1}\left(\frac{i}{n}\right).$$

How do you expect will this approximately look like?

6.35 A population consists of three numbers, namely 1, 1, and 5. Draw with replacement two numbers from this population.

a Determine the population mean μ and population variance σ^2.

b Derive the complete pdf of the sample.

c Calculate for each possible sample the sample mean as well as the sample variance. Use these results and your answer for part b to derive successively $\mathrm{E}(\overline{X})$, $\mathrm{Var}(\overline{X})$ and $\mathrm{E}(S^2)$. Do the results correspond with theory?

6.36 Take four random (but different) numbers (these four numbers form a population), put them in a spreadsheet, and determine the population mean and population variance. Now, suppose that two numbers are drawn with replacement from this population. Note that there are 4x4=16 different samples possible, each with probability 1/16. Enumerate all 16 samples and calculate for each sample the sample mean as well as the sample variance. Note that the complete distribution of the sample mean as well as the sample variance can be derived!

a Calculate the variance of the sample mean. Does it correspond with Theorem 6.8?

b Calculate the expected value of the sample variance. Does it correspond with Theorem 6.10?

6.37 Repeat problem 6.36, but now without replacement.

6.38 (B&E 8.13) Consider a sample ($n = 16$) from the standard normal distribution.

a Find $P(\overline{Z} < 0.5)$ b Find $P(Z_1 - Z_2 < 2)$ c Find $P(Z_1 + Z_2 < 2)$

d Find $P\left[\sum_{i=1}^{16} Z_i^2 < 32\right]$. e Find $P\left[\sum_{i=1}^{16}\left(Z_i - \overline{Z}\right)^2 < 25\right]$.

6.39 Determine a 98% prediction interval for the sample mean of a sample from a normal distribution with $n = 81$, and $\mu = 4.1$ and $\sigma = 1.3$.

6.40 One takes a sample of size $n = 12$ from a normal distribution with $\sigma^2 = 100$. Determine the probability that the sample variance will be between 50 and 240.

6.41 Consider a random variable X with a $N(500, 4^2)$-distribution.

a Determine a 95%-prediction interval for X, i.e. find I for which $P(X \in I) = 95\%$.

b Determine also a 95%-prediction interval for the sample mean of seven observations.

c Determine a 95%-prediction interval for the sample variance of seven observations.

d Determine a 95%-prediction interval for $\sum(X_i - \overline{X})^2$ of seven observations.

e Determine a 95%-prediction interval for $\sum(X_i - 500)^2$ of seven observations.

f (Computer) Simulate 200 samples of size $n = 7$ from the given distribution. Calculate for each sample the sample mean and the sample variance. Check whether the results correspond with the answers for b and c.

6.42 (B&E 8.15) Suppose that $X_i \sim N(\mu, \sigma^2)$, $i = 1,\ldots,n$ and $Z_i \sim N(0, 1)$, $i = 1,\ldots,k$, and all variables are mutually independent. State the distribution of each of the following random variables if it is a "named" distribution or otherwise state "unknown".

a $X_1 - X_2$ b $X_2 + 2X_3$ c $\dfrac{X_1 - X_2}{\sigma S_z \sqrt{2}}$ d Z_1^2

e $\dfrac{\sqrt{n}(\overline{X} - \mu)}{\sigma S_Z}$ f $Z_1^2 + Z_2^2$ g $Z_1^2 - Z_2^2$ h $\dfrac{Z_1}{\sqrt{Z_2^2}}$

i $\dfrac{Z_1^2}{Z_2^2}$ j $\dfrac{Z_1}{Z_2}$ k $\dfrac{\overline{X}}{\overline{Z}}$ l $\dfrac{\sqrt{nk}(\overline{X} - \mu)}{\sigma\sqrt{\sum_{i=1}^{k} Z_i^2}}$

m $\dfrac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma^2}+\sum_{i=1}^{k}(Z_i-\bar{Z})^2$ n $\dfrac{\bar{X}}{\sigma^2}+\dfrac{\sum_{i=1}^{k}Z_i}{k}$

o $k\bar{Z}^2$ p $\dfrac{(k-1)\sum_{i=1}^{n}(X_i-\bar{X})^2}{(n-1)\sigma^2\sum_{i=1}^{k}(Z_i-\bar{Z})^2}$

6.43 Derive the variance of the sample variance for samples of size n from a normal distribution with parameters μ and σ^2.

6.44 (B&E 8.17) Use tabled values to find the following: (Remark: for some of the probabilities you need the tables from Bain and Engelhardt, Appendix C, as sometimes the tables in the reader are too limited).

a $P(7.26 < Y < 22.31)$ als $Y \sim \chi^2(15)$

b The value b such that $P(Y < b) = 0.75$ if $Y \sim \chi^2(23)$

c $P\left[\dfrac{Y}{1+Y} > \dfrac{11}{16}\right]$ if $Y \sim \chi^2(6)$

d $P(0.87 < T < 2.65)$ if $T \sim t(13)$

e The value b such that $P(T < b) = 0.60$ if $T \sim t(26)$

f The value c such that $P(|T| > c) = 0.02$ if $T \sim t(23)$

g $P(2.91 < X < 5.52)$ if $X \sim F(7,12)$

h $P(1/X > 0.25)$ if $X \sim F(20,8)$

6.45 Let $X_1,\ldots,X_{10}$ be a sample from an exponential distribution, $X_i \sim \text{EXP}(\lambda)$.

a Determine $P\left[\dfrac{X_1+X_2+X_3}{X_1+X_2+X_3+X_4+X_5+X_6+X_7+X_8+X_9} < \dfrac{1}{2}\right]$.

b Derive the pdf of $W = \dfrac{X_1+X_3}{X_2+X_4+X_6+X_8+X_{10}}$.

6.46 (B&E 6.28) Let X_1 and X_2 be a random sample of size $n = 2$ from a continuous distribution with pdf of the form $f(x) = 2x$ for $0 < x < 1$ and 0 otherwise.

a Find the pdf's of the smallest and largest order statistic $Y_1 = X_{1:2}$ and $Y_2 = X_{2:2}$.

b Find the simultaneous pdf of Y_1 en Y_2.

c Find the pdf of the sample range $R = Y_2 - Y_1$.

6.47 Let $X_1,\ldots, X_n$ be a sample of size n from a continuous distribution with pdf $f(x) = 4x^3$ for $0 < x < 1$.

a Find the pdf of $Y_1 = X_{1:n}$ (the smallest order statistic).

b Give the simultaneous pdf of (Y_1, Y_2) with $Y_1 = X_{1:n}$ and $Y_2 = X_{2:n}$

c Are Y_1 and Y_2 independent? Explain why/why not.

6.48 (B&E 6.29) Consider a random sample $X_1,\ldots, X_n$ from a distribution with pdf $f(x) = 1/x^2$ for $1 < x$.

a Give the simultaneous pdf of all order statistics.

b Give the pdf's of the smallest and the largest order statistic, $Y_1 = X_{1:n}$ and $Y_n = X_{n:n}$.

c Derive the pdf of the sample range, $R = Y_n - Y_1$ for $n = 2$.

d Give the pdf of the sample median, Y_r, assuming that n is odd so that $r = (n + 1)/2$.

6.49 (B&E 6.30) Consider a random sample $X_1,\ldots, X_5$ ($n = 5$) from a Pareto(1, 2)-distribution with pdf $f(x) = \dfrac{2}{(1+x)^3}$ for $0 < x$.

a Give the simultaneous pdf of the second and fourth order statistics, $Y_2 = X_{2:5}$ and $Y_4 = X_{4:5}$.

b Give the simultaneous pdf of the three order statistics, Y_1, Y_2 en Y_3.

c Give the CDF of the sample median, Y_3.

6.50 (B&E 6.31) Consider a random sample $X_1, \ldots, X_n$ from an exponential distribution, $X_i \sim \mathrm{EXP}(1)$.

a Give the pdf of the smallest and largest order statistic, $Y_1 = X_{1:n}$ and $Y_n = X_{n:n}$.

b Give the pdf of the sample range, $R = Y_n - Y_1$.

c Give the simultaneous pdf of the first r order statistics, $Y_1, \ldots, Y_r$.

6.51 (B&E 6.33) Consider a random sample $X_1, \ldots, X_n$ from a geometric distribution, $X_i \sim \mathrm{GEO}(p)$. Give the CDF of the following: the minimum, Y_1, the kth smallest, Y_k and of the maximum Y_n.

7 Estimation

7.1 Introduction

We will leave now probability theory behind us and move on to *(inferential) statistics.* We no longer assume the (distribution of the) population to be known; instead, we will try to draw conclusions about these unknowns, based on one or more observed random samples. In order to be able to derive the proper techniques, the theory which has been developed in the past chapter is essential. The two main ways of drawing conclusions are parameter estimation and hypothesis testing.

The next course, Prob. Theory and Statistics 3, is dedicated to a number of very important concepts within inferential statistics. But like most books on Mathematical Statistics, the way of presenting this material in the book by Bain and Engelhardt may be too abstract for people who are not yet acquainted to estimation theory and hypothesis testing. Bain and Engelhardt present a general mathematical theory which may be used in the derivation of the correct results, *given a certain type of population distribution*. For each type of population distribution (e.g. Gamma, normal, uniform, or specified by a specific pdf) it is possible to derive dedicated estimators for the involved parameters (like θ or *r* for the Gamma distribution). Similarly, hypothesis tests may depend heavily on the type of distribution. For a novice student in this field, the underlying concepts may easily be drowned in the mathematical discussion. Therefore, we limit ourselves here to the estimation and the testing of general characteristics of a population, like its mean μ, the variance σ^2 and the proportion *p.* And if we assume a certain distribution of the population, it will only be the normal approximation. This chapter will focus on elementary estimation theory, while the next chapter will discuss basic hypothesis testing.

7.2 Point estimators

> ***Definition 7.1*** (*B&E, Def. 9.1.1*)
>
> A sample statistic which is used to estimate a (function of a) population parameter is called an **estimator** or **point estimator** (Dutch: schatter of puntschatter) for the parameter. The observed value (outcome, realisation) of the estimator based on a specific sample is called an **estimate** or a **point estimate** of the parameter.

Below, we will see that the sample mean is an estimator for the population mean. Note that an estimator is a random variable; the value the estimator will assume depends on the actual sample taken. So an estimator will have a certain distribution (usually unknown, since we no longer assume the population to be known). On the contrary, the estimate is just a constant, found after observing a specific sample.

7.2.2 The population mean μ

If the population mean μ is unknown and needs to be estimated, what is then more straightforward then using the sample mean $\overline{X}$ for that purpose? In Theorem 6.8 we have seen already that:

$$E(\overline{X}) = \mu$$

Thus, the sample mean possesses a very nice property: its expected value is equal to μ. We say that the sample mean provides an **unbiased estimator** (Dutch: zuivere schatter) for the population mean. Informally: if we would take an infinite number of random samples, then the mean of all those sample means will be equal to the population mean (unless the population variance is infinite, but that will only happen in rather uncommon situations).
However, an unbiased estimator does not necessarily mean that it is also 'the best' estimator. Consider for example an (artificial) estimator for the mean height of adult Dutch males which estimates their mean height in 50% of all samples as being 82 cm, and in another 50% of all sample as 282 cm. If the true mean is 182 cm, then actually this estimator is indeed unbiased, but clearly not of much help,

because we will always be 100 cm wrong! So another desirable property of an estimator is that the observed value of the estimator should lie relatively close to the true (but unknown) parameter value, in this case the population mean. That will be the case when the variance of an unbiased estimator is relatively small. When we have the choice between two estimators, both unbiased, then we should choose the estimator with the smallest variance (we will call that estimator more 'efficient' than the other; these concepts will be further discussed in the next course).

The *variance of the sample mean* is determined in Theorem 6.8 (under the requirement that we are dealing with a sample from a distribution):

$$\sigma_{\bar{X}}^2 = \mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}$$

The *standard deviation of the sample mean* is:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Note that this variance (and the standard deviation) becomes smaller for larger sample sizes. This is a confirmation of the intuitive result that sample means based on large samples will be more close to each other than sample means based on small samples. Because the sample mean is an unbiased estimator, this means that the observed values of the sample mean are more concentrated around the population mean than those for smaller samples.

Without proof, we give here the formula for the variance of the mean in a sample *without replacement* of size n from a finite population of size N:

$$\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}\frac{N-n}{N-1}$$

Note that we have encountered the factor $\dfrac{N-n}{N-1}$ earlier in the variance of the hypergeometric distribution. This factor is known under the name *finite-population-correction-factor*.

7.2.3 The variance σ^2

We will estimate the population variance by the sample variance (see Definition 6.6):

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Of course, there should be a reason for the remarkable fact that the denominator in the estimator above is equal to the sample size *minus* 1. In Theorem 6.10, we have seen already that in this way, we obtain an unbiased estimator of the population variance, since

$$\mathrm{E}(S^2) = \sigma^2$$

Without the '*minus* 1', we would structurally underestimate the population variance! To see this in another way: first assume that we are asked to estimate the population variance in the (rather uncommon) situation where the population mean μ is known. In that case, the sample variance can be determined by the formula $\frac{1}{n}\sum(x_i - \mu)^2$ (without the '*minus* 1'), which can be simply shown to be an unbiased estimator of σ^2. But since μ is usually unknown as well when we try to estimate σ^2, we have to estimate μ by $\bar{x}$ in the sum of squares. However, it is easy to prove that the sum of squares $\Sigma(x_i - a)^2$ is minimised by $a = \bar{x}$ (see problem 7.3). This means that $\sum(x_i - \bar{x})^2$ is always less than or equal to $\sum(x_i - \mu)^2$! Therefore, it is logical that, after replacing μ by $\bar{x}$ in the formula $\frac{1}{n}\sum(x_i - \mu)^2$, we should replace the denominator n as well by a denominator that is smaller than n.

We are then comparing each x_i with the mean of all x_i's; knowing $\overline{x}$, only (n – 1) of the x_i's are 'free to vary', and we say that one degree of freedom was lost.

Remark. Although S^2 is an unbiased estimator for σ^2, this does not mean that S (the sample standard deviation) is an unbiased estimator for σ (the population standard deviation), so in general:

$$\mathrm{E}(S) \neq \sigma$$

(this can also be seen from Jensen's inequality, see Theorem 3.7: $\mathrm{E}(S) \leq \sqrt{\mathrm{E}(S^2)}$). It is not possible to derive a general result for an unbiased estimator for σ, because it depends on the actual population distribution. Fortunately, the degree of bias is usually not very large, so often we will use S to estimate σ nevertheless.

Remark. When we consider a random sample from a finite population, then in general the result $\mathrm{E}(S^2) = \sigma^2$ does not hold. However, when the population is large, the bias will be quite limited.

7.2.4 The population proportion

Say we want to estimate the proportion of Dutch adults wearing glasses, which is exactly equal to the probability that an arbitrarily selected Dutch adult wears glasses. Since it is impossible to investigate all members in the population, we will try to estimate this population proportion p (or probability) using a random sample. In this sample, we will determine the proportion of people wearing glasses. The number of successes X in the random sample will follow a binomial distribution. The sample proportion is then (see also section 6.4.3):

$$\hat{P} = \frac{X}{n}$$

In Theorem 6.11, it has already been shown that:

$$\mathrm{E}(\hat{P}) = p$$

So the sample proportion is indeed an unbiased estimator of the population proportion. We also know:

$$\mathrm{Var}(\hat{P}) = \frac{pq}{n}$$

which shows that the estimates will become more and more concentrated around p for larger and larger sample sizes.

Remark. In sample without replacement from a finite population, the variance becomes smaller:
$$\mathrm{Var}(\hat{P}) = \frac{pq}{n}\frac{N-n}{N-1}.$$

7.3 Confidence intervals (or interval estimators)

7.3.1 Introduction

Although the point estimates of the previous section are very useful, they also have one big shortcoming: they do not provide very clear information about the quality of the estimate. Of course, when we look at the variance, we will be able to say something about this quality, but even then it is still not always easy to interpret this.
When a sample of the height of adult Dutch males results in a sample mean of 180.17647058823529 cm, then everyone will understand that the 'true' mean height will not be exactly equal to this sample mean. But probably, it will be somewhere 'in the neighbourhood' of 180.2 cm. We would like to be able to conclude something like: we are 95% confident that the true mean height will be between 179.5 and 181.7 cm. These confidence intervals consist of a range of values that act as good estimates of the unknown population mean. However, the interval computed from a particular sample does not necessarily include the true value of the mean.

How frequently the observed interval contains the true population mean if the experiment is repeated is called the confidence level. In other words, if confidence intervals are constructed in separate experiments on the same population following the same process, the proportion of such intervals that contain the true value of the parameter will match the given confidence level. When we say, "we are 95% confident that the true value of the parameter is in our confidence interval", we mean that 95% of the hypothetically observed confidence intervals will hold the true value of the parameter. After any particular sample is taken, the population parameter is either in the interval, or it is not. Since the observed data are random samples from the true population, the confidence interval obtained from the data is also random, so both the lower and the upper limit of the interval are random variables prior to the observation of a specific sample.

Definition 7.2

Let $X_1,...,X_n$ be a sample from a distribution with an unknown parameter θ. If L and U are sample statistics such that:

$$P(\ L < \theta < U\ ) = \gamma$$

then the (random) interval (L, U) is a 100γ%-**confidence interval** (Dutch: betrouwbaarheids-interval) for θ with **confidence level** γ. (Dutch: betrouwbaarheid).
But also the observed interval (l, u) (based on a specific sample) is usually called a confidence interval. (The latter l stands for *lower limit,* and u for *upper limit*)

Remark. The letter l in the definition above stands for *lower limit,* and u for *upper limit*.

Although it may seem more complicated, we will often use the notation $100(1-\alpha)$%-confidence intervals, where α = 1 − γ. If a corresponding hypothesis test is performed (see next chapter), α represents the level of significance (Dutch: onbetrouwbaarheid).

For the point estimators in the previous section, it was not important to know the type of population distribution. But in the derivation of confidence intervals, we do need to have knowledge about the type of distribution. The confidence intervals for μ and σ^2 which will be discussed below are valid only for samples taken from normal distributions. But because of the Central Limit Theorem, some of the results will still hold for samples from other distributions, as long as the sample size is large enough.

7.3.2 Confidence interval for μ (normal distribution, σ known)

We start by deriving a confidence interval for μ using the rather artificial presumption that the population standard deviation σ is known. The requirement that σ is known will be dropped in a later section.

Example 7.1

An instrument measures the degree of pollution of a soil sample with a measurement error which is normally distributed with an expected value of 0 and a standard deviation of 3 mg/m^3, independent of the degree of actual soil sample. Consider the pollution in the sample to be homogeneous over the sample, so if different values are recorded at different measurements for the same soil sample, then this is due to the measurement error of the instrument.
Now for a certain soil sample, four measurements are taken, with the following results:

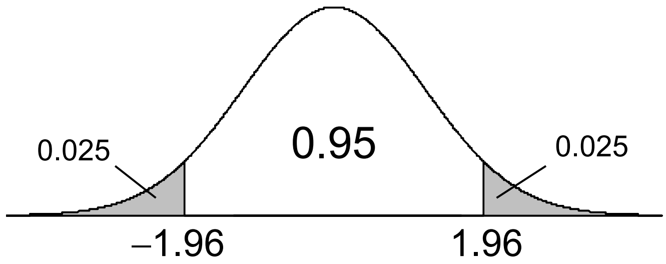
147.1 146.0 149.9 152.6

The true but unknown level of pollution in the sample, μ, is the expected value for each measurement, since the expected value of the measurement error is 0 here. This means that the measurements X follow a normal distribution with expectation μ and standard deviation 3 (all in mg/m^3).

Now, the sample mean $\bar{X}$ will also be normally distributed (Theorem 6.12), with expected value μ, and variance $\sigma^2/n = 9/4$, so $\bar{X} \sim N(\mu, 9/4)$ or $\dfrac{\bar{X}-\mu}{\sqrt{9/4}} = \dfrac{\bar{X}-\mu}{1.5} \sim N(0,1)$.

In order to derive a 95%-confidence interval for μ, we start by noting that ($z_{0.025}$ = 1.96):

$$\Rightarrow \quad P(-1.96 < \frac{\bar{X}-\mu}{1.5} < 1.96) = 0.95$$



The previous probabilistic statement should now be rewritten such that μ ends in the middle of two inequalities. This results in:

$$P(\ \bar{X} - 1.96(1.5) < \mu < \bar{X} + 1.96(1.5)\ ) = 0.95$$

The interval $(\bar{X} - 2.94,\ \bar{X} + 2.94)$ is then the 95%-confidence interval for μ. Since the sample above has a sample mean $\bar{x}$ = 148.9, we arrive at the observed interval (145.96, 151.84). ◄

We will repeat this procedure, but now in general terms. For samples from a normal distribution, we know that $\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,\ 1)$. This random variable is a function of μ and of the sample mean $\bar{X}$. Note that the distribution of this random variable does *not* depend on μ, and therefore we are able to write:

$$P(-z_{\alpha/2} < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < z_{\alpha/2}) = 1-\alpha$$

In section 6.4.4 , we have already discussed a $(1-\alpha)$100%-*prediction interval* for $\bar{X}$ as:

$$(\mu - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \mu + z_{\alpha/2}\frac{\sigma}{\sqrt{n}})$$

This form is only useful when we know μ already and want to predict the possible values for the sample mean. But now we want to estimate μ using an observed sample mean. So we rewrite the probability statement such that μ ends up in the middle of two inequalities, and we get (check):

$$P(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1-\alpha$$

Thus, the two confidence limits L and U (from Definition 7.2) for μ are found, and the 100(1–α)%-confidence interval for μ is:

$$(\ \bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\ ,\ \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\ )$$

Note the essential difference between the confidence interval for μ and the prediction interval for $\bar{X}$! But since they have both been derived from the very same formula, we can also say: the confidence interval is the set of those values of μ which would result in prediction intervals containing the observed sample mean.

Although the discussed procedure in finding confidence intervals is only correct when the sample was taken from a normal distribution, the Central Limit Theorem helps us in case of other population distributions as well, but only if the sample size n is large enough. As a rule of thumb it is often said that n should be at least 30 (but this is absolutely no guarantee that the result is approximatively correct).

The length (or width) of the 100(1−α)%-confidence interval for μ is the distance between the lower and the upper limit, so $2\cdot z_{\alpha/2}\,\sigma/\sqrt{n}$. The half of this width, so $z_{\alpha/2}\,\sigma/\sqrt{n}$ is usually called the **margin of error** (Dutch: schattingsfout of onnauwkeurigheidsmarge) at a confidence level of 1 – α. If it is necessary to determine the minimal value of n such that the margin of error is less than a prescribed value b, then we should solve n from the inequality $z_{\alpha/2}\,\sigma/\sqrt{n} \le b$.

Example 7.2

As a follow-up of Example 7.1, we are now interested in the question how large a sample size should be in order to make sure that the margin of error of a 90%-confidence interval for the degree of pollution should not exceed 0.5 mg/m^3. Because it is given that $\sigma = 3$, we will solve n from:

$$z_{0.05}\,\sigma/\sqrt{n} \le 0.5 \quad\Rightarrow\quad 1.645\cdot 3/\sqrt{n} \le 0.5 \quad\Rightarrow\quad \sqrt{n} \ge 1.645\cdot 3/0.5 \quad\Rightarrow\quad n \ge 97.42$$

Of course, n should be integer-valued, so the minimal sample size is 98. ◄

It is also possible to derive one-sided confidence intervals. Above, we started by writing:

$$P(-z_{\alpha/2} < \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < z_{\alpha/2}) = \gamma = 1-\alpha$$

But if we replace $z_{\alpha/2}$ by –∞, then an alternative starting point is:

$$P(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < z_{\alpha}) = 1-\alpha$$

Check that this will result in the one-sided 100(1–α)%-confidence interval ($\overline{X} - z_\alpha \dfrac{\sigma}{\sqrt{n}}$, ∞). The limit on the left-hand side is called the 100(1–α)%-confidence lower limit. In a similar way, we can also find (one-sided) confidence upper limits.

Before we can discuss the situation where the population standard deviation is not yet known, we will discuss first the confidence interval for σ.

7.3.3 Confidence interval for σ (normal distribution)

In order to derive a confidence interval for σ or σ^2, we need a function of sample observations which is also a function of σ, but not of any other unknown parameters. Furthermore, it should be possible to find the distribution of this random variable, where the distribution itself does not depend on any unknown parameters. If we manage to do so, then we will also be able to find a lower and an upper limit such that this random variable will fall within these limits with a probability of 100(1 – α)%. This interval can then be rewritten, such that the confidence interval for σ or σ^2 results.

In our search for the suitable random variable, note first that $\sum_{i=1}^{n}\frac{(X_i-\mu)^2}{\sigma^2} \sim \chi^2(n)$ (see Theorem 6.13). If we would know the true value of μ, then we can use this random variable to determine a confidence interval for σ or σ^2. However, almost always, the value of μ will be unknown, and thus we cannot use this random variable. But if we now replace μ in the formula above by the sample mean, then the distribution of the random variable changes as well. This is the situation discussed in Theorem 6.14, which is one of the most important theorems in statistics:

$$\frac{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2}{\sigma^2}=\frac{(n-1)S^2}{\sigma^2}\sim\chi^2(n-1)$$

This is exactly the type of situation we need: $\frac{(n-1)S^2}{\sigma^2}$ is a function of only the unknown variable σ , but with a distribution which does not depend on σ. Thus, we can write:

$$P(\ \chi^2_{n-1;\,1-\alpha/2}<\frac{(n-1)S^2}{\sigma^2}<\chi^2_{n-1;\,\alpha/2})=1-\alpha$$

The above formula can be rewritten as (check!!!)

$$P(\frac{(n-1)S^2}{\chi^2_{n-1;\,\alpha/2}}<\sigma^2<\frac{(n-1)S^2}{\chi^2_{n-1;\,1-\alpha/2}})=1-\alpha$$

So the $(1-\alpha)100\%$-confidence interval for σ^2 is:

$$\left(\frac{(n-1)S^2}{\chi^2_{n-1;\,\alpha/2}}\ ,\ \frac{(n-1)S^2}{\chi^2_{n-1;\,1-\alpha/2}}\right)$$

A confidence interval for σ can then be found from the confidence interval for σ^2 by simply taking the square roots of both limits.

The confidence interval above is only valid for samples taken from normal distributions. And while the Central Limit Theorem could come to our rescue for large samples when we discussed confidence intervals for μ in the previous section, that is no longer the case here. Even for large values of n , the quality of the interval remains not very good, unless the sample is taken from a distribution which is approximately normally distributed (at least symmetrical and unimodal).

Remark. We have seen that $\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2$ follows a chi-square distribution with n degrees of freedom. Here, each observation X_i is compared with the population mean μ. Because each observation is independent from all other observations, we say that n is the degrees of freedom. But after replacing μ by $\bar{X}$, we compare in $\sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{\sigma}\right)^2$ each observation with the mean of all observations. This introduces a dependency: given a value for $\bar{x}$, we only need to know the values of (n – 1) observations in order to be able to calculate the last observation: $x_n=n\bar{x}-\sum_{i=1}^{n-1}x_i$ (check!). We say that in the comparisons of X_i with $\bar{X}$, one degree of freedom is lost, resulting in n – 1 remaining degrees of freedom.

Remark. With the notation $\chi^2_{n-1;\,\alpha}$, we mean here the right-tail critical value such that $P(\chi^2(n-1)>\chi^2_{n-1;\alpha})=\alpha$. This notation is similar to the notation z_α for the standard normal distribution, and likewise for the student t-distribution. Although this is a very common notation, other authors, like Bain & Engelhardt, use the left-tail critical values; they write $\chi^2_{n-1;\,1-\alpha}$ where we write $\chi^2_{n-1;\,\alpha}$. Be aware of these differences!

7.3.4 Confidence interval for μ (normal distribution, σ unknown)

In section 7.3.2 , we treated the standard deviation to be known, but that is rarely the case. Usually, only the n observations from the sample are available, which can be used to determine the sample mean and sample standard deviation. As a starting point for the derivation of a confidence interval for μ, we can no longer use the random variable $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\sim N(0,\ 1)$, but we need to replace σ by its estimator, the sample standard deviation S, such that $\frac{\bar{X}-\mu}{S/\sqrt{n}}$ results. However, this last random variable is not normally distributed; fortunately, the solution lies in the next theorem:

Theorem 7.1 (*B&E, Th.8.4.3*)

If $X_1, X_2, \ldots, X_n$ is a sample from a normal distribution, with $X_i \sim N(\mu,\sigma^2)$, then:

$$T_{n-1} = \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$$ (so a t-distribution with n-1 degrees of freedom, see § 6.3)

<u>*Proof*</u>

$$T_{n-1} = \frac{\bar{X}-\mu}{S/\sqrt{n}} = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\frac{\sigma}{S} = \frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\frac{S}{\sigma}} = \frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{S^2}{\sigma^2}}} = \frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S^2}{\sigma^2}\Big/(n-1)}}$$

In the numerator of the last expression above, we recognise a standard normal random variable, while the denominator is the square root of a χ^2-distributed random variable divided by its number of degrees of freedom. Also, the numerator is independent of the denominator, because of Theorem 6.14. So this is exactly the situation as described in Theorem 6.4, and T_{n-1} has indeed a t-distribution with n-1 degrees of freedom.

Therefore, we can write the prediction interval for $\frac{\bar{X}-\mu}{S/\sqrt{n}}$ as:

$$P(-t_{n-1;\alpha/2} < \frac{\bar{X}-\mu}{S/\sqrt{n}} < t_{n-1;\alpha/2}) = 1-\alpha$$

This can be rewritten to:

$$P(\bar{X} - t_{n-1;\alpha/2}\frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{n-1;\alpha/2}\frac{S}{\sqrt{n}}) = 1-\alpha$$

This results in the 100(1−α)%-confidence interval for μ:

$$\left(\bar{X} - t_{n-1;\alpha/2}\frac{S}{\sqrt{n}},\ \bar{X} + t_{n-1;\alpha/2}\frac{S}{\sqrt{n}}\right)$$

For increasing samples sizes, the value of S will be more and more concentrated around σ. We expect therefore that the difference between the confidence interval above and the confidence interval discussed in 7.3.2 will disappear for very large sample sizes. That is indeed the case; in section 6.3, we remarked already that the t-distribution converges to the standard normal distribution for increasing sample sizes.

<u>*Example 7.3*</u>

Consider a sample of 10 oranges from a batch of oranges, with the following observed weights (in grams):

127 136 164 176 124 146 155 167 107 168

The sample mean is 147 and the sample standard deviation is 22.82 (check). When we assume the weight distribution to be normally distributed, then the 95%-confidence interval for the mean weight is:

$$(147 - 2.262\cdot 22.82/\sqrt{10},\ 147\ +\ 2.262\cdot 22.82/\sqrt{10}) = (130.7,\ 163.3)$$

($t_{9;\alpha/2}$ = 2.262 in the t-distribution with nine degrees of freedom at ½α = 0.025; check this) ◄

Remark. Just as before, the confidence interval derived in this section is only valid for samples from normal distributions. But the quality of the interval is actually rather robust, meaning that for large values of n it usually does not matter much when sampling is done from other distribution instead. As a rule of thumb, we might say that we can use this confidence interval when n is at least 30, unless the distribution is really very different from a normal distribution.

7.3.5 Confidence interval for p *of a binomial distribution*

Confidence intervals for the unknown probability p of success (or the population proportion) can be found in a similar way. When we take a sample from a Bernoulli distribution with probability p of success, then the number of successes X has a binomial distribution, with expectation np and variance npq (see previous course). Recall that whenever n is large enough, we can use a normal approximation for the distribution of X. And when we divide X by n, we obtain the sample proportion $\hat{P}$, of which we can see that:

$$\hat{P} \overset{approx}{\sim} N(p, \frac{p(1-p)}{n})$$

(This is a rather informal notation, meaning that $\hat{P}$ is *approximately* normally distributed when n is large enough; more formal notations will have to wait until the course Prob. Theory and Statistics 3).
When we know p, then the prediction interval for $\hat{P}$ follows from:

$$P(p - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} < \hat{P} < p + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}) \approx 1-\alpha$$

When we need to find a confidence interval for p, based on an observed sample proportion $\hat{P}$, we should rewrite the formula above until p ends up in the middle of two inequalities. Doing so actually turns out to be more complicated compared to the other confidence intervals, because we have to solve quadratic equations. Although that is definitely possible, the resulting confidence interval is not used often in text books. Instead, we will follow here the most commonly presented confidence interval for p, by first rewriting the above to (check!):

$$P(\hat{P} - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} < p < \hat{P} + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}) \approx 1-\alpha$$

However, the limits for p are not workable yet, because on both sides we still can see the unknown value of p. The idea is now that when we simply replace p in those limits by $\hat{P}$, we are introducing an extra error in the quality of the confidence interval, but an error which is fortunately in general only quite small. We obtain:

$$P(\hat{P} - z_{\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}} < p < \hat{P} + z_{\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}) \approx 1-\alpha$$

The most commonly used $100(1-\alpha)\%$-confidence interval for p is thus:

$$\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\ ,\ \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

So how large should n be in order to use the normal approximation to the binomial distribution? An often used rule of thumb says that n should be such that $np \geq 5$ and $nq \geq 5$, so for a value of p around 0.5, a value of n = 10 should be enough; for very small or very large values of p, we need a much larger value of n. Since p is unknown, we use as the rule of thumb: $n\hat{p} \geq 5$ and $n(1-\hat{p}) \geq 5$.

Like we discussed at the end of section 7.3.2 , we can determine the value of n such that the resulting margin of error will not be larger than a prescribed value b at a confidence level of $(1-\alpha)$. But the margin of error is here $z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, which can only be calculated after the sample is taken. This problem can be met in two different ways. Firstly, it is possible that we have already a reasonable estimate for p, maybe because of a smaller *pilot* sample. If so, we can use that estimate in the determination of n. If not, then note that the maximal value of the margin of error is determined by the product $\hat{p}(1-\hat{p})$. Since the maximum of this product is ¼ (reached when $\hat{p} = 0.5$, check!), we can find the necessary sample size from the inequality $z_{\alpha/2}\sqrt{\frac{1/4}{n}} \le b$. In this way, we will always be on the safe side: if after observing the sample it becomes clear that $\hat{p}$ is not close to 0.5, then the true margin of error will always be *smaller* than b. Thus, by taking the product to be equal to ¼, we really consider the worst case.

7.4 Confidence intervals for differences between two populations

7.4.1 Independent samples versus paired samples

Regularly we are interested in comparing two different populations. For example, is there a difference between two machines, both filling bags of potatoes, with respect to the mean filling weights? Is there a difference between the performance of boys and girls for the entrance test in elementary school? Is there a difference in the mean lifespans of two different brands of laptops? Is there a difference between the pain experience of patients before and after they have completed a particular therapy?
Although at first sight these examples look very similar, they can nevertheless represent two fundamentally different situations. First of all, we need to determine whether we are dealing with two independent samples, so are the random variables X_1, ... ,X_n of the sample from the first population independent from the random variables Y_1, ... ,Y_m which represent the sample from the second population? In the first of the three examples above, that is very likely the case; boys and girls are actually members of different populations and if we used a proper random sampling technique, then the resulting samples will also be independent. But the fourth example needs extra attention: if it is the case that *the same group* of patients were used before and after following a certain therapy, then we are no longer dealing with independent samples, but with a so-called *paired sample*. For each patient, we will have one pair of observations. These two different situations should also be treated statistically in quite different ways. With paired samples, it is usually possible to determine the difference between the two observations of each pair. These differences can then be treated as a sample (from one single observation), which means that we can employ the confidence intervals discussed earlier in this chapter.

Example 7.4

In an experiment, 80 patients with a chronical form of headache are asked to assign numbers from 0 (no headache) to 10 (most extreme form of headache). Each patient is asked twice for such an assessment, where X_i represents the score of patient i in a week where a new drug was given to the patient, and Y_i represents the score of patient i in a week where just a placebo was given. Neither the patient, nor the doctor seeing the patient knows which week the placebo or the drug was given (double blind research). In this way 160 scores are obtained, but in order to say anything about the effectiveness of the drug in comparison to the placebo, the statistical analysis will focus on the differences $X_i - Y_i$. These differences form a sample from a population with a sample size of 80, and we can use a technique like discussed in section 7.3.4 . ◄

If it is possible to set up an experiment using paired samples instead of independent samples, that may be beneficial, since it might reduce the variance in the samples. This can be illustrated by the following example:

A manufacturer of car tyres wants to compare the average lifespans of two different types of tyre. The manufacturer may choose one out of the two following experimental designs:

I Choose at random 40 cars, and fix 20 of those cars with tyres of type A, and the other 20 with tyres of type B.
II Choose 20 cars, and fix those cars with tyres of type A on one side and with tyres of type B on the other side (as determined by chance).

In both cases, the cars are driven by their respective owners. In the first case, we are dealing with two independent samples, while the second case uses a paired sample. The drawback for the first design is that all kind of factors linked to individual driving habits (fast/slow acceleration, brisk braking, etc.) cause a variation in the lifespans which may easily be much larger than the variation caused by the difference between tyres of type A versus type B. This may make it very difficult to draw any conclusion about this latter difference. In the second design, those factors contribute equally to the lifespans of both types of tyres, such that a difference in lifespans will result in an observation which will be much less sensitive to these individual differences in driving habits.

For the rest of this section, we will focus of the situation with two *independent* samples taken from normal distributions, since the paired sample case does not need any further theoretical exploration.

7.4.2 Confidence interval for $\mu_1 - \mu_2$ (known σ_1 and σ_2)

(B&E, pages 403)

We start here with the assumption that the standard deviations of the two populations are known. One of these two populations will be labelled as population 1, and the other as population 2. We will now develop an interval estimator for the difference between the two population means, so for $\mu_1 - \mu_2$. The most obvious choice for a *point* estimator for $\mu_1 - \mu_2$ is the difference between the two sample means, so $\bar{X}_1 - \bar{X}_2$. It is easy to see that $\bar{X}_1 - \bar{X}_2$ is an unbiased estimator for $\mu_1 - \mu_2$. And if the populations are normally distributed, or if both samples are large enough (CLT!), then we know that both sample means are normally distributed as well. From Theorem 4.13, we can derive that $\bar{X}_1 - \bar{X}_2$ should then also be normally distributed (but why? In Theorem 4.13 a plus sign is used, while here we are dealing with a minus sign. But we can also write $\bar{X}_1 - \bar{X}_2$ as $\bar{X}_1 + (-\bar{X}_2)$, where $(-\bar{X}_2)$ is again normally distributed with the same variance as $\bar{X}_2$). So:

$$\bar{X}_1 - \bar{X}_2 \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}) \quad \text{or} \quad \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

where n_1 and n_2 are the two samples sizes. Note the plus sign in the variance of this difference! This is all which is needed to find a suitable confidence interval, since from

$$P(-z_{\alpha/2} < \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} < z_{\alpha/2}) = 1 - \alpha$$

it follows that

$$P(\bar{X}_1 - \bar{X}_2 - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < (\mu_1 - \mu_2) < \bar{X}_1 - \bar{X}_2 + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}) = 1 - \alpha$$

The 100(1-α)%-confidence interval is therefore equal to:

$$(\bar{X}_1 - \bar{X}_2 - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} , \ \bar{X}_1 - \bar{X}_2 + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}})$$

This interval is exact for samples from two normal populations, and approximately correct for samples taken from other distributions, as long as both n_1 and n_2 are sufficiently large.

7.4.3 Confidence interval for $\mu_1 - \mu_2$ (unknown $\sigma_1 = \sigma_2$)

The situation in the previous section is rather rare; if we do not know the means, then we usually also do not know the standard deviations. In this section, we will assume only that the two standard deviations of the two populations are equal to each other (but unknown).

If $\sigma_1 = \sigma_2 = \sigma$, then we can write the random variable from the previous section as

$$\frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim \mathrm{N}(0,1)\,.$$

Before we can use this random variable to derive a confidence interval, note that we should replace σ by an estimator of σ. Both the sample variances S_1^2 and S_2^2 are in fact unbiased estimators of σ^2. We should somehow combine the information from both these estimators and come up with one single, new estimator for σ^2. We can do that by taking a weighted average of S_1^2 and S_2^2, and we will use the so-called *pooled sample variance*:

$$S_p^{\,2} \;=\; \frac{(n_1-1)S_1^{\,2} + (n_2-1)S_2^{\,2}}{n_1+n_2-2}$$

Since $\mathrm{E}(S_p^{\,2}) \;=\; \dfrac{(n_1-1)\mathrm{E}(S_1^{\,2}) + (n_2-1)\mathrm{E}(S_2^{\,2})}{n_1+n_2-2} = \dfrac{(n_1-1)\sigma^2 + (n_2-1)\sigma^2}{n_1+n_2-2} = \sigma^2$

it is clear that $S_p^{\,2}$ is an unbiased estimator of σ^2 as well.

Furthermore, it is simple to see that $\dfrac{(n_1+n_2-2)S_p^{\,2}}{\sigma^2} \sim \chi^2(n_1+n_2-2)$ (see problem 7.29).

In an analogous way as used in Theorem 7.1, we can now show that:

$$\frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{S_P^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{n_1+n_2-2}$$

This in turns leads to the following confidence interval:

$$\left(\overline{X}_1 - \overline{X}_2 - t_{n_1+n_2-2;\,\frac{1}{2}\alpha}\sqrt{S_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}\;,\quad \overline{X}_1 - \overline{X}_2 + t_{n_1+n_2-2;\,\frac{1}{2}\alpha}\sqrt{S_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}\right)$$

Example 7.5

Two machines are being used to fill bags with beans, and we want to know if there is any difference in the mean weights of the bags filled with the two machines. Since both machines are of the same brand, it seems reasonable to assume that they both do the filling with the same accuracy (= standard deviation), so that any differences in the mean weights can be attributed to a different calibration when tuning the machines for these particular bags. Two samples are taken:

Machine	Observations									n	$\overline{x}$	s^2
I	392	377	369	381	389	378	382	376	385	9	381	49.5
II	358	366	368	382	363	364	375			7	368	65

Since we assume that $\sigma_1 = \sigma_2 = \sigma$, we can estimate σ^2 by the pooled sample variance:

$$s_p^2 = \frac{(9-1)49.5 + (7-1)65}{9+7-2} = 56.14$$

The total number of observations is equal to 16, but two degrees of freedom were lost (one when calculating s_1^2 and the other when calculating s_2^2. So the remaining number of degrees of freedom is 16 – 2 =14, and the corresponding critical value for a 95% -confidence interval is equal to 2.145. Thus:

$$\left(381 - 368 - 2.145\sqrt{56.14\left(\frac{1}{9}+\frac{1}{7}\right)},\; 381 - 368 + 2.145\sqrt{56.14\left(\frac{1}{9}+\frac{1}{7}\right)}\right) = (4.901,\, 21.099)$$

This result shows us that we can be rather confident that a true difference exists between the mean filling weights of both machines. ◄

Note that the confidence interval in this section is only valid whenever the assumption of equality of variances is reasonable. We can however test this assumption separately (see section 8.5.2). That test can then determine whether we can use the confidence interval discussed here, or whether it is better to use the one discussed in the next section.

7.4.4 Confidence interval for $\mu_1 - \mu_2$ (unknown σ_1 and σ_2)

Now we will also drop the assumption of equality of the two standard deviations of the two populations. But then we immediately run into a problem, since we are not able to determine the exact distribution of $\dfrac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\left(\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}\right)}}$, which in turn means that an exact confidence interval does not exist!

However, several approximations have been developed over time which work quite well in many situations. The most common approximation method is by stating that the random variable above has approximately a student t-distribution, with an estimated number of degrees of freedom equal to

$$\nu = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{[(s_1^2/n_1)^2/(n_1-1)] + [(s_2^2/n_2)^2/(n_2-1)]}$$

This results is the following confidence interval:

$$\left(\bar{X}_1 - \bar{X}_2 - t_{\nu;\frac{1}{2}\alpha}\sqrt{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)} \;,\quad \bar{X}_1 - \bar{X}_2 + t_{\nu;\frac{1}{2}\alpha}\sqrt{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)} \right)$$

7.4.5 Confidence interval for σ_1^2/σ_2^2 (normal distribution)

(B&E, example 8.4.1)
When comparing the population means of two populations, we have used a confidence interval for the *difference* of the two means. That was successful, since we were able to find a random variable which is a function of $\mu_1 - \mu_2$ and of its estimator $\bar{X}_1 - \bar{X}_2$, and which has a distribution not involving $\mu_1 - \mu_2$. But for the difference $\sigma_1^2 - \sigma_2^2$ something similar is mathematically impossible. Instead, we will focus on the ratio σ_1^2/σ_2^2 instead of the difference $\sigma_1^2 - \sigma_2^2$, simply because we will see below that we will manage to derive a confidence interval for this ratio.

We know already that the sample variance is an unbiased estimator for the population variance. Now consider $\dfrac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$; if we are able to find the distribution of this random variable (and if this distribution does not involve σ_1^2/σ_2^2), then we can use this to derive a confidence interval for σ_1^2/σ_2^2 .

Theorem 7.2 *(B&E, Ex.8.4.1)*

If $X_1,...,X_{n_1}$ and $Y_1,...,Y_{n_2}$ are independent samples, with $X_i \sim N(\mu_1,\sigma_1^2)$ and $Y_i \sim N(\mu_2,\sigma_2^2)$, then:

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{S_1^2\sigma_2^2}{S_2^2\sigma_1^2} \sim F(n_1-1, n_2-1)$$

Proof

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{\frac{(n_1-1)S_1^2}{\sigma_1^2}\Big/(n_1-1)}{\frac{(n_2-1)S_2^2}{\sigma_2^2}\Big/(n_2-1)} \sim F(n_1-1, n_2-1),$$

This is because we know that $\frac{(n_i-1)S_i^2}{\sigma_i^2} \sim \chi^2(n_i-1)$ $(i=1,2)$ (see Theorem 6.14), such that the result follows directly from Theorem 6.6.

This theorem makes it possible to write the following probability statement:

$$P(f_{n_1-1,n_2-1;1-\alpha/2} < \frac{S_1^2\sigma_2^2}{S_2^2\sigma_1^2} < f_{n_1-1,n_2-1;\alpha/2}) = 1-\alpha$$

(here $f_{n_1,n_2;\alpha}$ denotes the right-tail critical value of an F-distribution)

This can now be rewritten to a probability statement with the ratio $\frac{\sigma_1^2}{\sigma_2^2}$ in the middle (check!):

$$P(\frac{S_1^2}{S_2^2}\frac{1}{f_{n_1-1,n_2-1;\alpha/2}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{S_1^2}{S_2^2}\frac{1}{f_{n_1-1,n_2-1;1-\alpha/2}}) = 1-\alpha$$

which gives us the 100(1 – α)%-confidence interval for the ratio of the two population variances.

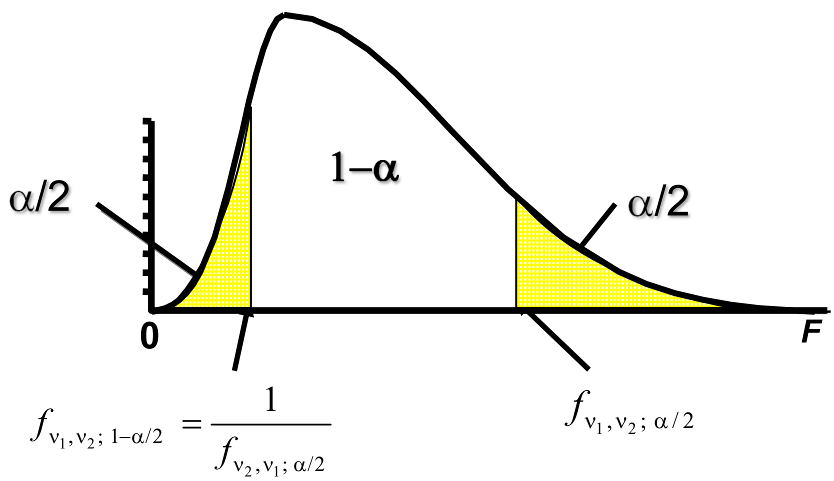
Earlier (see page 68) we saw that the left-tail critical values, like $f_{n_1-1,n_2-1;1-\alpha/2}$, is usually not shown in tables for the F-distribution. So (unless you live in modern times and can use a computer), you can use the equality

$$\frac{1}{f_{n_1-1,n_2-1;1-\alpha/2}} = f_{n_2-1,n_1-1;\alpha/2}\ .$$

Thus, we can write the confidence interval for σ_1^2/σ_2^2 as:

$$\left(\frac{S_1^2}{S_2^2}\frac{1}{f_{n_1-1,\,n_2-1;\,\alpha/2}}\ ,\quad \frac{S_1^2}{S_2^2}f_{n_2-1,\,n_1-1;\,\alpha/2}\right)$$

For the confidence interval of the ratio σ_1/σ_2, we only need to take the square roots of the interval limits above.



7.4.6 Confidence interval for $p_1 - p_2$

In order to derive a confidence interval for the difference between two population proportions, note first that we know from section 7.3.5 that $\hat{P} \overset{approx}{\sim} \mathrm{N}(p, \frac{p(1-p)}{n})$, so this will also be valid for the two sample proportions:

$$\hat{P}_1 \overset{approx}{\sim} \mathrm{N}(p_1, \frac{p_1(1-p_1)}{n_1}) \text{ and } \hat{P}_2 \overset{approx}{\sim} \mathrm{N}(p_2, \frac{p_2(1-p_2)}{n_2}).$$

Therefore, we get

$$\hat{P}_1 - \hat{P}_2 \overset{approx}{\sim} \mathrm{N}(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2})$$

Or: $$\frac{\hat{P}_1 - \hat{P}_2 - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \overset{approx}{\sim} \mathrm{N}(0,1)$$

Thus

$$P(-z_{\alpha/2} < \frac{\hat{P}_1 - \hat{P}_2 - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} < z_{\alpha/2}) \approx 1 - \alpha$$

which can be rewritten as:

$$P(\hat{P}_1 - \hat{P}_2 - z_{\alpha/2} \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} < (p_1 - p_2) < \hat{P}_1 - \hat{P}_2 + z_{\alpha/2} \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}) \approx 1 - \alpha$$

For the very same reason as discussed earlier in section 7.3.5 , this cannot be used yet as a confidence interval. Therefore, we replace the population proportions in the two limits by the sample proportions:

$$P(\hat{P}_1 - \hat{P}_2 - z_{\alpha/2} \sqrt{\frac{\hat{P}_1(1-\hat{P}_1)}{n_1} + \frac{\hat{P}_2(1-\hat{P}_2)}{n_2}} < (p_1 - p_2) < \hat{P}_1 - \hat{P}_2 + z_{\alpha/2} \sqrt{\frac{\hat{P}_1(1-\hat{P}_1)}{n_1} + \frac{\hat{P}_2(1-\hat{P}_2)}{n_2}}) \approx 1 - \alpha .$$

7.5 Problems

7.1 Given is a sample of size 12 with the observed values:

3 8 5 6 5 4 7 6 7 5 7 9

a If this is a sample from a population (e.g. grades for an exam) with mean μ and standard deviation σ, calculate the estimate of these μ and σ.

b Give an estimate of the standard deviation of the sample mean when the twelve observations are a sample from a probability distribution.

7.2 Take your pocket calculator and find out how to switch to Statistics Mode (e.g. mode SD) and how to enter a series of numbers only once, such that mean and both variances (for sample and population) can be calculated directly. Also, find out how the variances and the outcome of $\sum (X_i - \overline{X})^2$ can be determined in a simple manner.

7.3 Prove that $\sum_{i=1}^{n} (x_i - a)^2$ is minimal for $a = \overline{x}$. For this reason $\overline{x}$ is also known as the least-squares-estimator.

7.4 Consider a sample of heights (in cm) of army recruits collected during a physical test with observed heights:

165, 188, 183, 181, 178, 169, 194, 181, 176 and 175.

Calculate sample mean, sample variance and sample standard deviation. Calculate also the sample proportion of heights below 180 cm.

7.5 Consider a sample (n=3, (X_1, X_2, X_3)) from a distribution and the following estimators for the population mean: $X_A = \frac{1}{2}X_1 + \frac{1}{2}X_2$ $\quad X_B = \frac{1}{5}X_1 + \frac{2}{5}X_2 + \frac{2}{5}X_3$ $\quad X_C = \frac{2}{5}X_1 + \frac{2}{5}X_2 + \frac{2}{5}X_3$ and $X_D = \frac{1}{3}X_1 + \frac{1}{3}X_2 + \frac{1}{3}X_3$

a Which estimators are unbiased?

b Determine the variance of each of the unbiased estimators. Which estimator is the best?

7.6 This exercise is meant to show how it is possible to find the distribution for sample statistics completely in simple cases (of course only if the population is completely known). Next, from this distribution we can determine the bias and efficiency of the estimators.

Consider a box containing ten ones and ten fives.

a. Determine the population mean and the population variance.

b. A sample (n = 3) is drawn *with replacement* (X_1, X_2, X_3) from this box. Verify that in total 8 different samples can be drawn. Calculate for each of these samples the observed value of the following three sample statistics:

(i) $\overline{X}$: the sample mean

(ii) m: mediaan(X_1, X_2, X_3) (the middle observation in size)

(iii) S^2: the sample variance

Observed sample (x_1,x_2,x_3)	Probability of observed sample	$\overline{x}$	m	S^2
(1,1,1)				
(1,1,5)				

c. Derive from these results for each of the three sample statistics the probability distribution (in table format), and then determine the expected value and the variance for each sample statistic.

d. Check for each sample statistic whether they are unbiased estimators (and for which population parameters?). Which of the three is the best estimator for the population mean? Explain why.

7.7 A device measures the concentration of substance A in a liquid in micrograms per liter (μg/l). The random error of the device is normally distributed with a standard deviation of 7 μg/l. The observations from the same liquid (a certain amount of the liquid) are:

165, 188, 183, 181, 178, 169, 194, 181, 176, 175 (compare with exercise 7.4))

Determine the confidence interval for the concentration with a confidence level of

a	90%	b	95%
c	98%	d	99%
e	99.5%	f	99.9%

7.8 For a device for measuring the wavelength of light it is known that the standard deviation is 2.5 nanometer (1 nm = 10^{-9} m). The observed wavelengths of three measurements are:

567.4 570.0 568.1

a Determine a point estimate for the searched wavelength.
b Calculate the 99%-confidence interval for the searched wavelength. Which assumption has to be made?
c Determine a 99%-confidence interval for the searched wavelength, but now one that is not based on 'equal tails' (make your own choice). What happens with the length of the confidence interval?
d If the 99%-confidence interval (with equal tails) may have a maximal length of 1 nm, how large should the number of observations be?
e Determine a left one-sided 99%-confidence interval (i.e. a 99%-confidence interval lower limit) for the searched wavelength.

7.9 A device measures radioactivity in a certain cave. One wishes to measure the radiation in the cave with a confidence level of 95% and a margin of error of 0.5 units. So, the 95%-confidence interval may have a width of 1 unit. The device has a standard deviation of 1.7 units. How large should the sample size be in order to achieve the required margin of error?

7.10 For a sample of weights of jam jars (normal distributed) the observed weights of the contents are:

456 458 454 454 455 451 457

Determine a 95%-confidence interval for the standard deviation of the weight of the contents.

7.11 For five videotapes the playing time is determined: 366, 339, 364, 356 and 379 (minutes).
Assume that these measurements are a sample from a normal distribution and determine a 95%-confidence interval for the variance σ^2 of this distribution. Also, give a 95%-confidence interval for σ.

7.12 Determine a 99%-confidence interval lower limit (one-sided interval) for the variance of a normal distribution for which the following nine observations are known:

−3 6 −7 8 4 0 2 12 −8

7.13 A sample from a normal distribution resulted in a sample standard deviation of $s^2 = 42$. Find the sample size if the 95%-confidence interval for the variance turns out to have a width of 100.

7.14 Verify that $P(\frac{(n-1)S^2}{\chi^2_{n-1;\,\alpha/2}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{n-1;\,1-\alpha/2}}) = 1-\alpha$ follows from $P(\chi^2_{n-1;\,1-\alpha/2} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{n-1;\,\alpha/2}) = 1-\alpha$.

7.15 Derive from $\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 = \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma^2} \sim \chi^2_n$ (mentioned at the beginning of 7.3.3) a (formula for a) confidence interval for σ^2 (only valid when μ is known).

7.16 Consider a sample of size 8 from a normal distribution with observed values:

-3 0 5 -7 -1 2 3 1

a Determine an equal tailed two-sided 95% confidence interval for the population mean μ.
b Determine a one-sided 95%-confidence lower limit for μ.

7.17 Consider a sample of size 8 from a normal distribution with observed values:

100 105 107 108 112 101 106 108 103 109 113 98 92 109 116 104

a Determine a 95%-confidence interval for μ.
b If we assume that the standard deviation of the distribution is equal to 10, how large should the sample size be to get a 95%-confidence interval with a width of at most 2?

7.18 For packaging of food sometimes heavy aluminium foil is used. A batch of aluminium foil contains an unknown number of sheets, all similar in size and neatly stacked, and weighing 5135 gram all together. In total six sheets are being weighed and the observed weights (in g) are:

1.74 1.69 1.72 1.78 1.73 1.72

a Calculate a 95%-confidence interval for the average weight μ of the sheets in the batch, assuming that the weights are normally distributed.

b With the results from part a, derive a 95%-confidence interval for the number of sheets in the batch.

7.19 Based on the observations of a sample from a normal distribution with unknown expectation μ the following values are known:

$$\sum x_i^0 = 20 \qquad \sum x_i^1 = 2546 \qquad \sum x_i^2 = 351\,624$$

Calculate the mean and variance of this sample and determine the 95%-confidence interval for μ.

7.20 Determine an interval which will contain an observation from a Bin(30, ½)-random variable X with about 95% probability.

7.21 In a sample from a large population, 350 out of the 700 respondents preferred drink B.

a Determine a 95%-confidence interval for the fraction of the population that prefers the drink.

b The market researcher considers the width of the interval too wide. For which sample size will the 95%-confidence interval have a width of at most 0.02?

c Answer the same questions for the 99%-confidence interval.

7.22 A survey of 300 students (the sample) showed that 50 of them are considering doing another study. Determine a 95%-confidence interval for the proportion of all students (the population which is much larger than the sample) that considers to change their study.

7.23 In a random sample of size 1250 one found 325 people approving a particular regulation in the city. Determine a 95%-confidence interval for the fraction of the total population of the city approving this regulation.

7.24 One examines companies with more than 100 employees. In the US a sample of 496 employees contained 62 executives; in a similar survey in Japan, one found 30 executives in a sample of size 810. Determine the 95%-confidence intervals for the fraction of executives in the US and in Japan.

7.25 An investor can choose between two investment portfolios. The return of the first portfolio is normally distributed with (population) mean μ_1 and (population) standard deviation σ_1, while the return of the second portfolio is normally distributed with (population) mean μ_2 and (population) standard deviation σ_2. In the table below, the realised returns of these two portfolios are given for a random sample of 8 selected weeks from the year 1999.

Week	Portfolio 1	Portfolio 2
1	0.22	0.32
2	0.59	0.35
3	0.11	0.48
4	-0.05	0.36
5	0.44	0.27
6	0.38	0.33
7	0.05	0.47
8	0.00	-0.01

The investor wants an 80%-confidence interval for the difference in return of the portfolios. Construct a two-sided confidence interval for $\mu_1 - \mu_2$.

7.26 Two samples with each 13 observations from independent normal distributions show observed averages of 1005 and 1001 respectively. Also, for both distributions holds: $\sigma = 4$.

a Determine for both population means separately a 95%-confidence interval.

b Determine the 95%-confidence interval for $\mu_1 - \mu_2$.

7.27 For two branches A and B of company C the daily turnovers for a particular week (Mo / Sa) are given (x 1000 euro`s):

A: 151 174 200 131 168 316
B: 134 155 183 149 154 245

Calculate the pooled sample variance.

7.28 For two similar alloys K and L the solidification temperatures have been determined several times. The sample statistics are:

Alloy	Sample size	Mean	Standard deviation	Variance
K	4	512	12	144
L	7	534	15	225

a Calculate the pooled sample variance.

b Determine a 90%-confidence interval for the difference in solidification temperatures of the two alloys, assuming that the variances in the measurements are the same for both alloys, and that the measurements follow a normal distribution.

7.29 Which distribution has the following random variable: $\frac{(n_1+n_2-2)S_p^{\,2}}{\sigma^2} = \frac{(n_1-1)S_1^{\,2}+(n_2-1)S_2^{\,2}}{\sigma^2}$? (Both samples are from two populations with normal distribution with unknown, but equal variances). Use this result to proof that $\frac{\overline{X}_1 - \overline{X}_2 - (\mu_1-\mu_2)}{\sqrt{S_P^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$ has a t-distribution with n_1+n_2-2 degrees of freedom (see also 7.4.3).

7.30 Comparing the normally distributed burning time of two types of lamps the sample statistics are as follows:

	sample size	sample mean	sample variance
Lamp X	13	984	8742
Lamp Y	15	1121	9411

Determine the 95%-confidence interval for the difference of the expectations $\mu_X - \mu_Y$

a under the assumption $\sigma_X^2 = 9000$ and $\sigma_Y^2 = 9500$.

b under the assumption $\sigma_X^2 = \sigma_Y^2$, but with unknown value.

c without any assumption for σ_X and σ_Y.

7.31 Medical costs may be dependent on whether or not a seat belt is used while driving. A series of similar accidents is broken down by seat belt use. The following information is available on the medical costs:

	number	sample mean	sample standard deviation
With seatbelt (X)	15	565	220
Without seatbelt (Y)	12	1200	540

a First, assume $\sigma_X = 220$, $\sigma_Y = 540$ (which implies of course different information from the observed values above) and determine a 95%-confidence interval for $\mu_X - \mu_Y$. (Assume that the costs for each category are approximately normally distributed).

b Determine a similar confidence interval, now only using the assumption of equality between σ_X and σ_Y.

7.32 Two different machines for filling bottles of vinegar are compared. With both machines 10 bottles are filled with 1 liter of vinegar. In this case it comes to comparing the variances of the filling amounts. Assume that the deviations per machine are approximately normally distributed. The sample standard deviations for the machines are $s_1 = 0.123$ and $s_2 = 0.208$ (ml), respectively.

a Determine 95%-confidence intervals for σ_1 and σ_2 separately.

b Determine a 95%-confidence interval for the ratio σ_1 / σ_2 .

7.33 Ten of 25 surveyed male students economics believe that statistic is the most difficult course, while from 36 surveyed female students 21 have the same opinion. Construct a 95%-confidence interval for the difference between the percentages of students in the two populations (male and female) who have the opinion that statistics is the most difficult course.

7.34 A research is conducted on the percentage of families (with children) having two incomes. First a group of 950 families (group 1) is considered and 5 years later 500 families (group 2). It turned out that in the first group 664 families had two incomes, and in the second group 360. Give a 90%-confidence interval for the difference in the percentages of households with 2 incomes.

8 Hypothesis testing

8.1 Fundamental concepts

The hypotheses

The presumption that a coin is fair, or that second hand cars of two different brands are on average just as expensive, or that male and female employees in similar functions earn the same: these are all examples of hypotheses. In statistical hypothesis testing, a hypothesis states something about one or more populations. At first, we will assume a certain hypothesis to be true, followed by an analysis of one or more samples to see if there is proof which contradicts this hypothesis. The hypothesis which is used as our initial assumption is called the **null hypothesis** (H_0 , Dutch: nulhypothese).

If the null hypothesis is not true, then something else should be true. For example, male employees earn on average more than their female colleagues. Or: the coin is not fair. Such a statement is called the **alternative hypothesis** (H_a or H_1 , Dutch: alternatieve hypothese).

Criminal courts of justice provide us with an interesting analogue to statistical hypothesis testing. In our justice system, we assume that the accused is innocent (the null hypothesis), until the proof presented in court is considered to be strong enough (beyond reasonable doubt) to decide that the accused is guilty (the alternative hypothesis). Although a different model (guilty, until sufficient evidence exists of innocence) could be considered, we have chosen the ‘innocent until proven guilty’-principle: most people consider the error of convicting an innocent person as more serious as the error of acquitting a guilty person. When we use the terminology of statistical hypothesis testing, then we could say that the more serious first error is an **error of type I** (Dutch: fout van de eerste soort of type I fout), and the second error as an **error of type II**. And what does it mean exactly when an accused person is acquitted in the Dutch courts of justice? We cannot say that there is 100% proof that the accused is innocent (because if such proof would exist, or surface somewhere in the course of a trial, then the case will be dropped and an acquittal is not even necessary anymore). So when someone is acquitted, it merely means that there is insufficient proof that the accused is guilty, *and therefore* we hold on to our *assumption* of innocence.

With statistical hypothesis testing, there are always two possible conclusions: the null hypothesis is rejected (meaning there is ‘sufficient evidence’ for the alternative hypothesis) or the null hypothesis is not rejected (meaning there is insufficient evidence for the alternative hypothesis). So note that we can never prove the validity of the null hypothesis; if the purpose of hypothesis testing is to try to prove a particular statement, then that statement should be the alternative hypothesis! (That is why the alternative hypothesis is also called the 'research hypothesis').

The diagram below shows when a conclusion is correct and when it is not. We consider an error of type I as more serious than an error of type II. That is why we want to make sure that the probability of type I error will never be larger than a certain pre-agreed value α, which we will call the **significance level** (Dutch: significantieniveau, onbetrouwbaarheidsdrempel). Very often, a value of $\alpha = 0.05$ or smaller is chosen for the significance level. Thus P (reject H_0 | H_0 is true) $\leq \alpha$. The probability of a type II error is then subsequently a result of this chosen value: in general, decreasing the significance level will lead to an increase in the probability of an error of type II and vice versa.

Possible conclusions	Actual situation H_0 **is true**	 H_0 **is not true** (and H_a is true)
H_0 **is not rejected**	Correct conclusion	Error of type II
H_0 **is rejected**	Error of type 1	Correct conclusion

Example 8.1

A large home-painter company always orders its paint cans from the paint manufacturer Hiskens. The company starts doubting whether Hiskens fills the cans according to the nominal content of 750 ml. But the company wants to be quite sure before daring to risk the good relation it has with Hiskens. Thus, the null hypothesis is that Hiskens fills the cans with (on average) 750 ml of paint, and the alternative hypothesis that Hiskens fills those on average with less than 750 ml. So: H_0: $\mu = 750$ versus H_a: $\mu < 750$. ◄

The test statistic

The next step in hypothesis testing is to identify a sample statistic which provides information relevant to the question whether the null hypothesis should be rejected or not. Depending on the null hypothesis this can be, for example, the sample mean or the number of heads appearing when a coin is thrown a number of times. Such a statistic will be called a **test statistic** (Dutch: toetsingsgrootheid). Important: a test statistic can only be useful if its distribution can be determined *under the assumption that the null hypothesis is true.* We say that the distribution is known 'under H_0', or 'given the null hypothesis'. Knowing this distribution, we are able to distinguish values for this test statistic which are 'likely' from those which are 'unlikely', both using the assumption that the null hypothesis is true. If we then determine the observed value for the test statistic and this value belongs to the set of 'unlikely' outcomes, then we will reject the null hypothesis.

Example 8.2

The company from the previous example wants to test the hypotheses H_0: $\mu = 750$ versus H_a: $\mu < 750$. They decide to use a sample of 25 cans, where the contents each of the 25 cans will be measured accurately. It seems straightforward that the sample mean is a candidate for the choice of test statistic, since it clearly carries information concerning the possible correctness of the hypotheses. Everyone will intuitively agree that a sample mean very close to 750, or larger, will form no reason to reject H_0, while a sample mean which is much less could very well be a good reason to reject H_0. But the sample mean is only useful as a test statistic here if we are able to find its distribution under the assumption that H_0 is true. If we assume now that the population variance σ^2 is known, and that the contents of the cans are normally distributed, then we are indeed able to find the distribution under H_0, i.e. $\bar{X} \sim \mathrm{N}(750,\ \sigma^2/25)$ (see Theorem 6.12).

(Note that if σ^2 is not known, the distribution cannot be determined and the sample mean in itself is not a proper test statistic….) ◄

Example 8.3

At a selection procedure for darts players, two dart players A and B throw 25 times a dart at a specific spot on a board. After each throw of A and B, the coach checks whose dart is closest to the spot. Of these 25 throws, A wins 17 times and B only 8. The conclusion of the coach is that A is better than B, but B objects that this difference is not significant enough, and that this difference may have been caused by pure chance instead of a structural difference in the abilities of both players in hitting the spot. So let us start with the assumption (= null hypothesis) that both players are equally good. If that would be the case, then the probability of winning should be 0.50 for both players. If we now write p as the proportion of times A will win (in a potentially infinite series of attempts), then we can write the hypotheses as H_0: $p = 0.5$ versus H_a: $p \neq 0.5$.

We can choose here the *number* of wins for A in the series of 25 throws as our test statistic. This is a useful choice, since it clearly tells us something about the hypotheses. Furthermore, we are indeed able to find the distribution of this test statistic under the assumption that the null hypothesis is true, i.e. under the assumption that both players are equally good: its distribution is binomial with parameters n=25, p=½. Then the expected number of wins of A (again under H_0) is equal to 12½. The observed outcome 17 is definitely higher, but that in itself does not mean yet that this can be called an 'unlikely large' outcome….. ◄

The rejection region

So how can we determine which outcomes for the test statistic are so 'unlikely' under H_0 that they should lead to the rejection of H_0? The subset of the outcome space containing all those 'unlikely' outcomes is called the **rejection region** or **critical region** (Dutch: kritieke gebied, verwerpings-gebied). Whenever a sample shows an observed value for the test statistic which falls within the rejection region, the null hypothesis will be rejected. Recall that we want to be sure that the probability of a type I error is never larger than the chosen level of significance α, so P (reject H_0 | H_0 is true) $\leq$ α. This defines how large the rejection region can be, since

$$P(\text{test statistic falls within rejection region} \mid H_0 \text{ is true}) = P(\text{reject } H_0 \mid H_0 \text{ is true}) \leq \alpha.$$

This does not rule out the possibility that maybe many different rejection regions may exist, all satisfying the requirement above. During the next course, rejection regions will be developed using certain principles which may look a bit complicated at first. However, in this course common sense will be sufficient! Of course, when observing an 'unlikely' outcome, it is useful to reject H_0 only when that outcome will be more 'likely' under the alternative hypothesis. So the rejection region should consist of those values for the test statistic which are unlikely under H_0, and more likely under H_a.

Example 8.4

Consider again our painters company and the hypotheses H_0: $\mu = 750$ versus H_a: $\mu < 750$ using a sample of 25 cans. Now assume that the contents of the population of cans is normally distributed with $\sigma^2 = 400$. The test statistic with distribution under H_0 is $\overline{X} \sim N(750,\ 400/25)$. Our common sense tells us that, even though for example an observed mean of 900 is truly very unlikely under H_0, such an outcome should never lead to a rejection of H_0, since this outcome is even more unlikely when the alternative hypothesis would be true. Indeed, only relatively low values for the sample mean can be a reason to reject H_0, since low values will be more likely under H_a than under H_0. But how low should the sample mean be before we can reject the null hypothesis? Suppose that the company chooses a level of significance of α = 0.05, so the probability that the company will falsely accuse Hiskens is limited to a maximum of 5%. Let us now start by looking whether it is possible to choose as rejection region the set $\{\overline{x} \mid \overline{x} \leq 745\}$ (so reject H_0 whenever the sample mean is less than or equal to 745. Then:

$$P(\text{type I error}) = P(\text{ Test statistic in rejection region} \mid H_0 \text{ is true }) =$$
$$= P(\overline{X} \leq 745 \mid \mu = 750) = P(\frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \leq \frac{745 - 750}{20 / \sqrt{25}}) = P(Z \leq -1.25) = 0.1056$$

So the rejection region $\{\overline{x} \mid \overline{x} \leq 745\}$ would lead to a probability of a type I error which is too large. Apparently, the rejection region should be smaller. The correct determination of the rejection region can be performed as follows: we know that under H_0 that $Z = \dfrac{\overline{X} - 750}{20/\sqrt{25}} \sim N(0,\ 1)$ and since for the standard normal distribution we also know that $P(Z \leq -1.645) = 0.05\ (= \alpha)$, we get:

$$P(\frac{\overline{X} - 750}{20 / \sqrt{25}} \leq -1.645) = 0.05 \quad \Rightarrow \quad P(\overline{X} \leq 750 - 1.645 * 20 / \sqrt{25}) = 0.05$$
$$\Rightarrow \quad P(\overline{X} \leq 743.42) = 0.05$$

So the correct rejection region here is $\{\overline{x} \mid \overline{x} \leq 743.42\}$, because it leads to an acceptable probability of a type I error.

So what happens if we would make this rejection region even smaller, for example to $\{\overline{x} \mid \overline{x} \leq 740\}$? The probability of a type I error would then become smaller as well; however, that is not automatically better, because we will see later on that the probability of a type II error will *increase* as a result! Therefore, we choose the largest possible rejection region such that

$$P(\text{type I error}) = P(\text{ Test statistic in rejection region} \mid H_0 \text{ is true }) \leq \alpha.$$ ◄

One-tailed or two-tailed tests

The test in the last example is a so-called left-tailed (or left-sided; Dutch: linkséénzijdige) test. We were specifically interested to find proof that μ is less than a certain value μ_0 (the value mentioned in

H_0). But in other cases, we are interested to find proof that μ is different from μ_0, and we are dealing with a two-tailed test.

Example 8.5

A company has just received a very large order of steel bars from its supplier. The bars are supposed to have a diameter of 20mm. Assume that it is known (and accepted) that the diameters are normally distributed and that the standard deviation of the diameters is equal to 0.5mm. The company would like to know if the mean diameter of the bars is different from 20mm, so if $\mu \neq 20$. To that end, a random sample of 30 bars is taken, which will be used test the hypotheses $H_0 : \mu = \mu_0$ (=20) versus $H_a : \mu \neq \mu_0$ (=20). The test statistic will again be the sample mean $\overline{X}$ which has, given H_0, a normal distribution with mean 20 and a standard deviation $0.5/\sqrt{30}$. Because the alternative hypothesis is now two-sided, we should reject H_0 not only for a relatively small value of the sample mean, but also for a relatively large value! Given the null hypothesis, we know that:

$$P(-z_{\alpha/2} < \frac{\overline{X}-20}{0.5/\sqrt{30}} < z_{\alpha/2}) = 1-\alpha$$

$$\Rightarrow \quad P(20 - z_{\alpha/2}\, 0.5/\sqrt{30} < \overline{X} < 20 + z_{\alpha/2}\, 0.5/\sqrt{30}) = 1-\alpha$$

At a level of significance α of 5%, we obtain the following two-sided rejection region:

$$(-\infty,\ 20 - 1.96\frac{0.5}{\sqrt{30}}\,] \cup [20 + 1.96\frac{0.5}{\sqrt{30}},\ \infty) \quad \text{or} \quad (-\infty,\ 19.821] \cup [20.179,\ \infty).$$

(the probability of a type I error is for this rejection region exactly equal to $\alpha = 0.05$). Suppose now that the sample of 30 bars results in an observed sample mean of 20.16 mm. That value does not, given this rejection region, provide sufficient proof to state that the population mean is different from 20 mm. ◄

Whether we should test one-sided or two-sided depends on what we would like to know about the population, the research hypothesis (= alternative hypothesis). If, for example, we want to know whether the population mean is larger than some value, we arrive at a right-tailed test with H_0: $\mu = \mu_0$ against H_a: $\mu > \mu_0$.

Example 8.6

- Suppose the objective is to compare the effectiveness of two different teaching methods using exam results. If the interest is purely scientific, then a two-tailed test is called for.
- An educational institution is considering the introduction of a new, more expensive teaching method. They want to find out whether the new method leads to better (=higher) exam results. Only in that case, they may decide to adapt this new method. The test should be right-tailed; it is not interesting to know whether there is proof the new method may even give worse results.
- A researcher wants to determine whether lung cancer is more prevalent in regions with a high degree of air pollution. He chooses to use a right-tailed test, since it is theoretically inconceivable that a high degree of air pollution results in *less* lung cancer cases. ◄

An important consequence of the choice for testing one-tailed instead of two-tailed is that the rejection region will lie (on that side) closer to the value of μ_0 than in the two-tailed case. This can potentially lead to situations where the null hypothesis would be rejected when we choose a one-tailed test, while it would not be rejected if we would have tested two-sided.

Example 8.7 (Continuation of Example 8.5)

Consider Example 8.5 again, but now we will use a right-tailed test, so $H_a : \mu > \mu_0$ (=20). At the same value of $\alpha = 0.05$, we find the rejection region (check!):

$$[20 + 1.645\frac{0.5}{\sqrt{30}},\ \infty) \quad \text{or} \quad [20.150,\ \infty)$$

Again assume now that the sample mean results in the value 20.16. This means that we can now reject the null hypothesis, while we were not able to do so at a two-tailed test! ◄

It is important to stress that the choice of testing either one- or two-sided should be made *before* analysing any specific random sample. In general: there should be other reasons (theoretical, practical, historical,…) for choosing a one-sided test; this choice can never be based on the sample observations!

Rejection regions for test statistics with discrete distributions

In all examples above, we were able to find a rejection region such that the probability of a type I error is exactly equal to the significance level α. That is because we have been dealing with continuous type distributions for the test statistics. But in case the test statistic has a discrete distribution, this equality can seldom be reached, and we have to settle for the largest rejection region such that

$$P(\text{type I error}) = P(\text{ Test statistic in rejection region} \mid H_0 \text{ is true }) \leq \alpha.$$

Example 8.8 (Continuation of Example 8.3)

We consider here Example 8.3 again, but we will first discuss a *right-tailed test*, so H_0: $p = 0.5$ versus H_a: $p > 0.5$. The test statistic X is defined as the number of wins for A. Under H_0, X is binomially distributed, with $n = 25$ and $p = ½$. If we choose $\alpha = 0.05$, we will now try to find the correct rejection region for this test statistic. It should be clear that H_0 can only be rejected when the value of X is relatively large, so the rejection region therefore should have the form $\{x \mid x \geq c\}$, where the critical value c is chosen such that the probability of a error of type I is not larger than α (=0.05). Let us use a trial-and-error method to find the correct value for c, and we start with $c = 16$. Then we can use a table for the binomial distribution to find:

$$P(X \text{ falls in the rejection region} \mid H_0) = P(X \geq 16 \mid H_0) = 0.115.$$

So the rejection region $\{x \mid x \geq 16\}$ is not correct, since the probability of a type I error would be too large. Similarly, we can also easily see that the rejection region $\{x \mid x \geq 17\}$ is still (just) too large, but that the rejection region $\{x \mid x \geq 18\}$ suffices. The associated probability of a type I error (0.0216) is actually a lot smaller than 0.05, but because of the discrete character of the distribution of the test statistic it is impossible to enlarge the rejection region just a little. So if A wins 17 times, there is insufficient evidence to state that player A is better than B. ◄

Example 8.9 (Continuation of Example 8.3)

Now, we will consider the two-sided test H_0: $p = ½$, versus H_1: $p \neq ½$. The rejection region will consist now of the union of two parts: $[0, c_l] \cup [c_r, 25]$, where c_l is the left-tailed critical value and c_r is the right-tailed critical value. We are now searching for those values of c_l and c_r such that:

$$P(X \text{ in the rejection region} \mid H_0) = P(X \leq c_l \mid H_0) + P(X \geq c_r \mid H_0) \leq 0.05$$

Although again there will be different possible choices for c_l and c_r, it seems reasonable to choose for symmetry, such that $P(X \leq c_l \mid H_0) = P(X \geq c_r \mid H_0)$.
This results in the rejection region for X : $[0, 7] \cup [18, 25]$, because

$$P(X \leq 7) + P(X \geq 18) = 0.0216 + 0.0216 = 0.0432$$

Check that a larger rejection region will result in a probability of a type I error which will exceed α. So if A wins 17 times, then this does not provide sufficient proof for stating that a difference exists between the two players. ◄

Composite null hypotheses

All null hypotheses discussed until now are so-called **simple** hypotheses, meaning that the value of the population parameter is completely determined by the value in H_0, like e.g. $\mu = 750$. In contrast, the alternative hypotheses were always **composite** hypotheses, since the population parameter is *not* completely determined in case the alternative hypothesis is true (e.g. $\mu < 750$). So what happens if we consider the possibility that the null hypothesis is also composite (say H_0: $\mu \geq 750$).

Example 8.10 (continuation of Example 8.4

Consider again the hypothesis test as described in Example 8.4, but now with the null hypothesis

H_0: $\mu \geq 750$ instead of H_0: $\mu = 750$. The alternative hypothesis is still H_a: $\mu < 750$. So the rejection region should clearly lie again somewhere in the left-tail of the distribution of the test statistic $\overline{X}$. Suppose now that we use the *same* rejection region as before, so $\{\overline{x} \mid \overline{x} \leq 743.42\}$. Now we get

$$P(\text{ error of type I }) = P(\text{ H}_0 \text{ is rejected} \mid \text{H}_0 \text{ true })$$

$$= P(\text{ test statistic in rejection region} \mid \mu \geq 750) = P(\overline{X} \leq 743.42 \mid \mu \geq 750)$$

Note that this probability can no longer be determined, because the true value for μ is now no longer known. It is clear that this probability will depend on the true value of μ, and will be different for different values of μ. However, it is also clear that $P(\overline{X} \leq 743.42 \mid \mu \geq 750)$ will be a decreasing function of μ: if μ increases, then it will become less and less likely that the sample mean will be less than or equal to 743.42. Thus, the largest value for the probability $P(\overline{X} \leq 743.42 \mid \mu \geq 750)$ will be found when $\mu = 750$. Earlier, we have stated that P(test statistic in rejection region | H_0) should never exceed the value of α. Since $\mu = 750$ can be seen as a kind of 'worst case' for the probability of a type I error, we can draw now the conclusion that the rejection region remains exactly the same as it was. So, it does not matter at all for the testing procedure whether we use H_0: $\mu \geq 750$ or H_0: $\mu = 750$! ◄

For all hypothesis tests *discussed in this course*, the same applies as in the previous example, so there is no need to treat the case of a composite H_0 any different from the case of a simple H_0.

The relation between a hypothesis test and the confidence interval

Consider a two-sided test with H_0: $\mu = \mu_0$. Note that (see e.g. Example 8.5) we started the derivation of the rejection region (when using a level of significance α) by writing:

$$P(-z_{\alpha/2} < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}) = 1 - \alpha .$$

And this has also been the starting point for the derivation of a 100(1–α)%-confidence interval for μ! It should therefore not be surprising that an important relationship exists between a rejection region and a confidence interval. If the null hypothesis in a test with H_0: $\mu = \mu_0$ versus H_a: $\mu \neq \mu_0$ will not be rejected at a level of significance α, then μ_0 will be within the 100(1-α)%-confidence interval (that could have been determined using the same random sample. And vice-versa: if μ_0 is in the 100(1-α)%-confidence interval for μ, then the null hypothesis H_0: $\mu = \mu_0$ will not be rejected (at significance level α). Thus: we can also use a confidence interval in order to determine whether a certain null hypothesis should be rejected or not.

Example 8.11 (Example 8.5 continued)

In Example 8.5 we tested $H_0 : \mu = \mu_0$ (=20) against $H_a : \mu \neq \mu_0$ (=20) . For $\alpha = 0.05$, the rejection region for $\overline{X}$ was $(-\infty, 19.821] \cup [20.179, \infty)$. For an observed sample mean of 20.16 mm, we had to draw the conclusion that H_0 cannot be rejected.

But we could also have used the value of 20.16 mm to determine a 95%-confidence interval, which would be (19.981, 20.339) (check!). Because $\mu_0 = 20$ is within this interval, there is no reason to doubt H_0. In other words: the conclusion is exactly the same! ◄

This relationship might result in the feeling that all of hypothesis testing is just a set of unnecessary procedures. However, some important aspects of testing theory are not highlighted yet. Moreover, a one-sided hypothesis test would require a one-sided confidence interval. And although this creates no theoretical problems, working with one-sided confidence intervals soon becomes very confusing (should we use a lower-limit confidence interval, or an upper-limit confidence interval).

Remark. Among novice students, often there is a misconception that the critical region and the confidence interval are each other's complement. Despite the fact that the formulas in some cases resemble each other, nothing is further from the truth! The rejection region presupposes knowledge of the population (through H_0) and says something about the probability of certain outcomes for a test statistic. In contrast, the confidence interval presupposes knowledge of the sample, and says something about the possible values the (unknown) population parameter may possibly have. Be well aware of these differences! (But it is true that the rejection region and the *prediction interval* are each other's complement.)

8.2 Probability of a type II error, and power

The purpose of any hypothesis test is above all: to discover whether the true value of the population parameter deviates from the value in the null hypothesis. The probability that the test statistic falls within the rejection region, given that the alternative hypothesis is true, is called the **power** (Dutch: onderscheidingsvermogen) of a test.
An error of type II occurs when H_0 is not rejected, given that H_a is true. The probability of a type II error (often denoted by β) is therefore always equal to 1 minus the power. The four different possible cases are shown below; in three of the four cases, the probability has a specific name.

Decision	Actual situation	
	H_0 is true	H_a is true
H_0 is not rejected	$\geq 1-\alpha$	Probability of a type II error (β)
H_0 is rejected	Probability of a type I error ($\leq \alpha$)	Power ($1-\beta$)

Note that in both columns the probabilities add up to 1.

After a rejection region has been determined, we can determine the probability of a type II error, since:

$$\beta = P(\text{ type II error }) = P(\text{ Test statistic not in rejection region} \mid H_a \text{ is true }).$$

But this probability can only be found if we know the distribution of the test statistic under H_a. And since H_a is usually composite, there are many 'alternative' parameter values, and for each of those values we can calculate another value for β.

Example 8.12 (Continuation of Example 8.4)
In Example 8.4, we found the rejection region $\{\bar{x} \mid \bar{x} \leq 743.42\}$. Let us determine the probability of a type II error for two, arbitrarily chosen, values for the population mean, say for μ = 740ml and for μ = 730ml (note that in both cases the alternative hypothesis is true). First for μ = 740ml:

$$\begin{aligned} P(\text{type II error }) &= P(\text{ Test statistic not in rejection region} \mid \mu = 740) \\ &= P(\bar{X} > 743.42 \mid \mu = 740) = P(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} > \frac{743.42-740}{20/\sqrt{25}}) \\ &= P(Z > 0.855) \approx 0.195 \end{aligned}$$

And for μ = 730ml:

$$P(\bar{X} > 743.42 \mid \mu = 730) = P(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} > \frac{743.42-730}{20/\sqrt{25}}) = P(Z > 3.37) \approx 0.0004 \qquad ◄$$

The probability of a type II error will always decrease when the population parameter becomes more and more different from its value according H_0. Note also that the 'true' value of the population parameter is (and remains) unknown; we can only calculate the probability of a type II error under the *assumption* that the population parameter has a certain value.

Similarly, the power of a test is in general a function of the population parameter, equal to 1 – β.

Although the power is defined as the probability of rejecting H_0 given that H_a is true, usually the **power-function *π*** is defined simply as the probability of rejecting H_0 as a function of *all* possible values of the population parameter (so not only limited to the set mentioned in H_a. So, if μ is the population parameter in question, then we can write Power(μ) = π(μ) = *P* (reject H_0 | μ).

Example 8.13 (Continuation of Example 8.5)

In Example 8.5 we considered a test $H_0 : \mu = 20$ against $H_a : \mu \neq 20$, where μ is the population mean of steel bars delivered. In a sample of size 30 and a level of significance of 5%, the rejection region found was:

$(-\infty, 19.821] \cup [20.179, \infty)$

Assume now that the true mean diameter of the steel bars is equal to 20.1; what is then the probability that H_0 will be rejected? Or, stated otherwise: what is the power of this test when $\mu = 20.1$?

Power(μ =20.1) = π(20.1) = P(Test statistic in rejection region | μ =20.1) =

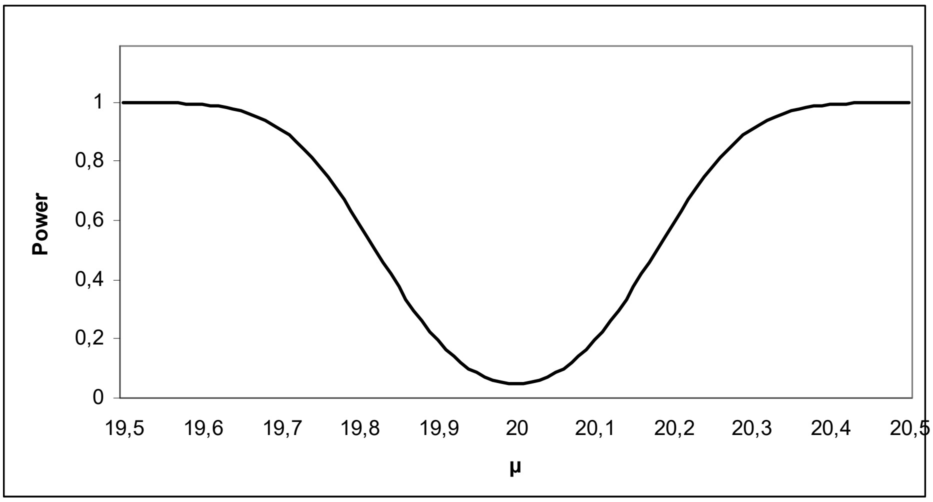
$$= P(\overline{X} < 19.821 \mid \mu = 20.1) + P(\overline{X} > 20.179 \mid \mu = 20.1)$$

$$= P(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < \frac{19.821 - 20.1}{0.5/\sqrt{30}}) + P(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} > \frac{20.179 - 20.1}{0.5/\sqrt{30}})$$

$$= P(Z < \frac{19.821 - 20.1}{0.5/\sqrt{30}}) + P(Z > \frac{20.179 - 20.1}{0.5/\sqrt{30}})$$

$$= P(Z < -3.06) + P(Z > 0.87) = 0.0011 + 0.1922 = 0.1933$$

In other words: the probability is only 19.3% that such a deviation of 0.1 mm in μ will be 'discovered' by this hypothesis test (with $n = 30$ and α=0.05).
If we would determine the power-function for other values of μ as well, the following figure results:



We see clearly that, in case the true value of μ is quite different from 20, the null hypothesis will be rejected with a large probability. The closer μ will be to 20, the smaller this probability becomes. If $\mu = 20$, then this probability is (of course!) exactly equal to the value of α, in this case 0.05. This figure shows the typical form of the power function belonging to a two-sided test.
Note that for a left-tailed test, the power function will converge to 0 for values of μ greater than 20. ◄

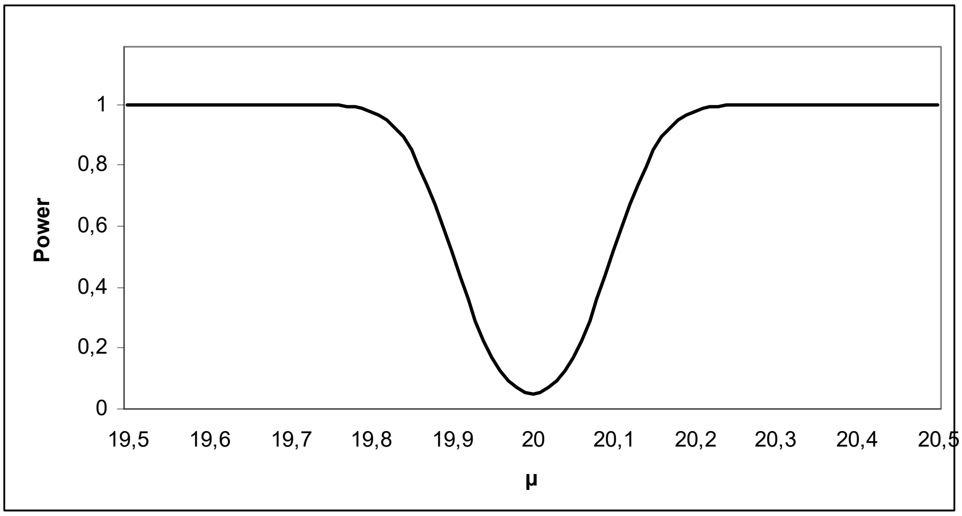
The power of a particular test depends directly on the choice of the significance level. If this level is reduced, the rejection region will need to be smaller, so that the power decreases as well. However, it is often possible to improve the power at a constant level of significance, namely by increasing the sample size. Normally this will lead to an increase in the size of the rejection region, so that the power will increase as well.

Example 8.14 (Continuation of Example 8.13)

When we increase the sample size in the previous example to $n = 100$, the rejection region becomes:

$(-\infty, 19.902] \cup [20.098, \infty)$ (check!)

This in turn leads to the following power function:



When we compare this power function with the power function in Example 8.13, we see that the power is now higher for all values of μ which appear in the alternative hypothesis. ◄

If we have the choice between two different tests, each with the same value of α, then it is obvious that we should choose the test which has the highest power for all values of the parameter as specified in the alternative hypothesis. This concept of 'Most powerful test' will be further elaborated during the course Prob. Theory and Statistics 3.

Example 8.15 (Example 8.13 continued)

Let us look again at the previous example with the steel bars. Now suppose someone suggests the following test: draw a marble from a box with 100 marbles, numbered from 1 to 100. If the number on the selected marble is 5 or less, then reject the null hypothesis $H_0 : \mu = 20$. If not, then do not reject the null hypothesis. What is the probability of a type I error?

Answer: $P(\text{ Reject } H_0 \mid H_0 \text{ is true }) = 5/100 = 0.05$

So this test meets the requirement regarding to the maximum acceptable probability of a type I error. Nevertheless, this test is rather absurd, because in no way we are using information about the diameters of the bars in a sample. So why is this test no good? The real problem of this test is the power, because the power is 0.05 *everywhere*, and as such much worse than the tests using the sample mean. ◄

8.3 The p-value

If in a scientific publication the conclusion is drawn that a certain null hypothesis is rejected at a specific value of α, then the reader can ask himself whether the same conclusion would have been possible for a smaller value of α. There is a large degree of arbitrariness in the selection of α, and simply reporting whether the null hypothesis is rejected tells us not a great deal about the strength of the evidence against H_0. Therefore, many publications show the so-called p-value.

The **p-value** (of a certain observed value for a test statistic in a sample) is the probability of obtaining a result equal to or "more extreme" than what was actually observed, assuming the null hypothesis is true. For example, if the observed value for the test statistic for a left-tailed test is x, then the p-value is equal to $P(\text{ Test statistic} \le x \mid H_0)$.

The p-value indicates how extreme the observed value for the test statistic really is in the distribution for the test statistic under the null hypothesis. The smaller the p-value, the more extreme the outcome, and therefore the less likely the outcome could occur if the null hypothesis is true. In this sense, a small p-value tells us that the null hypothesis may not adequately explain the observation, and can be seen as a measure of the "credibility" of the null hypothesis.

The level of significance α is the maximum probability of a type I error which, in a particular situation, is considered to be acceptable, so $\alpha \geq P(\text{ H}_0 \text{ is rejected} \mid \text{H}_0 \text{ is true })$. If the p-value is smaller than α, then H_0 should be rejected. Thus, we could also have defined the p-value as follows:

> The p-value for an observed sample outcome is equal to the smallest possible value of α at which the null hypothesis can still (just) be rejected.

Or stated the other way around: the rejection region consists exactly of those possible values for the test statistic which would result in a p-value less than or equal to α.

Example 8.16 *(continuation of Example 8.4)*

Consider again the hypothesis test as described in Example 8.4 with H_0: $\mu = 750$ versus H_a: $\mu < 750$, with $\alpha = 0.05$. Assume that we find a sample mean $\bar{x} = 742$ ($n = 25$). Thus, the p-value is simply:

$$\text{p-value} = P(\bar{X} \leq 742 \mid \mu = 750)$$

Since $\sigma^2 = 400$, this probability is easily found to be:

$$P(\bar{X} \leq 742 \mid \mu = 750) = P(\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \leq \frac{742 - 750}{20 / \sqrt{25}}) = P(Z \leq -2.0) = 0.0228$$

Since this probability is less than the chosen value for α (= 0.05), we reject H_0. Note that we would have arrived at the same conclusion when we compare $\bar{x} = 742$ with the rejection region found earlier ($\{\bar{x} \mid \bar{x} \leq 743.42\}$). ◄

In case a test has *a two-tailed* rejection region, we know that the test statistic *TS* might have an observed value in any of the two tails. If H_0 is true, then the probability of observing a value for the *TS* in the left-tailed part of the rejection region should never be larger than α *divided by two*. Assume now that we observe a value x somewhere in the left tail. Now, instead of comparing $P(\ TS \leq x \mid H_0\ )$ with $\alpha/2$, we will multiply $P(\ TS \leq x \mid H_0\ )$ by 2 and compare this result with α. This is also according to definition of the p-value above, since the observed outcome could just as likely have been somewhere in the other tail, both just as 'extreme'. In general, we can for two-tailed test write:

$$\text{p-value} = 2 * \min(P(\ TS \leq x \mid H_0\ ), P(\ TS \geq x \mid H_0\ )\ )$$

Note that logic tells us usually immediately which of the two probabilities will be the smallest, which one will be 'more extreme', simply by checking whether the outcome is somewhere in the left- or in the right tail.

Example 8.17 *(Example 8.5 continued)*

Consider Example 8.5 about the steel bars, with $H_0: \mu = \mu_0$ (=20) versus $H_a: \mu \neq \mu_0$ (=20). The observed sample mean was 20.16. This outcome is already in the right-tail of the distribution of the test statistic under H_0. When calculating the p-value, we therefore take the probability that the sample mean exceeds 20.16, multiplied by 2 (two-tailed test!).

$$\text{p-value} = 2 \cdot P(\bar{X} \geq 20.16) = 2 \cdot P(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \geq \frac{20.16 - 20}{0.5/\sqrt{30}}) = 2 \cdot P(Z \geq 1.75) = 0.0802\ .$$

So the probability of observing such a value (or more extreme) is equal to 8%, given the null hypothesis. Since this probability is larger than the chosen value of the level of significance, the null hypothesis should not be rejected. ◄

If the test statistic *TS* has (under H_0) a distribution which is symmetrical around 0, then the p-value for a two-tailed test can be written as well as (check!):

$$\text{p-value} = 2 * P(\ TS \ \geq\ |\,x\,| \ \mid H_0\ )$$

Remark. In Dutch textbooks, the p-value is often called the 'overschrijdingskans van de uitkomst' but in the case of a two-tailed test, it is actually twice either the 'linker- of rechteroverschrijdingskans'.

In English language, the p-value is regularly also called the "observed significance level". In computer output, it is often indicated simply as "significance". In a package like SPSS, this is usually the two-sided p-value, which means that in case we are actually working with a one-tailed alternative hypothesis, this value should be divided by two! Be aware that the "level of significance" is selected by the researcher before looking at sample data, while the “significance” is the p-value calculated after observing a sample. The conclusion by the researcher then depends on the comparison of these two values.

Unfortunately, the p-value is often misunderstood. Below we address a list of frequently held, but incorrect, beliefs about the p-value
- The p-value is *not* the probability that the null hypothesis is true. (The calculation of the p-value is based on the assumption that H_0 is true, so p-value = P (outcome | H_0) $\neq$ P (H_0 | outcome | H_0)).
- The p-value is *not* the probability that the null hypothesis is falsely rejected.
- 1 - (p-value) is *not* the probability that the alternative hypothesis is true.
- The level of significance of a test is *not* determined by the p-value.
- The p-value does *not* say anything about the size or importance of the measured effect.

8.4 Hypotheses concerning a single population

8.4.1 The hypothesis $\mu = \mu_0$ (normal distribution with known σ)

This is the case we have been dealing with already in the previous sections. Let us derive the rejection region once more for the two-tailed test:

$$H_0:\ \mu = \mu_0 \text{ versus } H_a:\ \mu \neq \mu_0$$

The sample mean $\bar{X}$ is a suitable test statistic: we have seen that it is an (unbiased) estimator for μ, and we know the distribution of the sample mean under H_0 (because we are still assuming the population variance to be known):

$$\bar{X} \sim N(\mu_0,\ \sigma^2/n)\text{, and so}$$

$$\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,\ 1)\,.$$

Using this, we can write (again, still assuming that H_0 is true):

$$P(-z_{\alpha/2} < \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} < z_{\alpha/2}) = 1 - \alpha$$

$$\Rightarrow\quad P(\mu_0 - z_{\alpha/2}\,\sigma/\sqrt{n} < \bar{X} < \mu_0 + z_{\alpha/2}\,\sigma/\sqrt{n}) = 1 - \alpha$$

The interval $(\mu_0 - z_{\alpha/2}\,\sigma/\sqrt{n},\ \mu_0 + z_{\alpha/2}\,\sigma/\sqrt{n})$ is sometimes called the 100(1-α)%-prediction interval for the sample mean (see also page 68). The complement of this interval is formed by the values of the test statistic which are rather unlikely under H_0, so that is the rejection region:

$$(-\infty\ ,\ \mu_0 - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\,]\ \cup\ [\mu_0\ +\ z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\ ,\ \infty)$$

With a right-sided alternative hypothesis H_a: $\mu > \mu_0$, the rejection region will be:

$$[\mu_0\ +\ z_{\alpha}\frac{\sigma}{\sqrt{n}}\,,\ \infty)$$

And for a left-sided test:

$$(-\infty\ ,\ \mu_0 - z_{\alpha}\frac{\sigma}{\sqrt{n}}\,]$$

Note that we could have chosen another test statistic here, which is the *standardised* sample mean

$$Z = \frac{\overline{X} - \mu_0}{\sigma / \sqrt{n}} \sim \mathrm{N}(0,1) \quad (\text{or } Z = \frac{\overline{X} - \mu_0}{\sigma}\sqrt{n} \sim \mathrm{N}(0,1)).$$

The two-tailed rejection region for this test statistic is:

$$(-\infty\,,\ -z_{\alpha/2}\,] \cup [\,z_{\alpha/2}\,,\infty)$$

The p-value at a specific sample mean $\overline{x}$ is:

for a right-tailed test: $P(\overline{X} \geq \overline{x}) = P\left(Z \geq \dfrac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}\right)$

for a left-tailed test: $P(\overline{X} \leq \overline{x}) = P\left(Z \leq \dfrac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}\right)$

for a two-tailed test: $2\min(P(\overline{X} \leq \overline{x}),\ P(\overline{X} \geq \overline{x})) = 2\,P\left(Z \geq \left|\dfrac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}\right|\right)$

The test discussed here is, due to the use of the standard normal distribution, often referred to as the z-test. If the population is not normally distributed, the Central Limit Theorem tells us that, if the sample size is sufficiently large, the sample mean is approximately normally distributed. That means that this test remains useful in those cases (rule of thumb: if $n \geq 30$).

8.4.2 The hypothesis $\mu = \mu_0$ (normal distribution with unknown σ)

Now we are dropping the rather uncommon assumption that σ is known. The two test statistics discussed in the previous section are now no longer of any use, since its distributions are not known under H_0. But we can estimate the population standard deviation σ by the sample standard deviation S. If the population is normally distributed, then (see section 7.3.4):

$$T_{n-1} = \frac{\overline{X} - \mu_0}{S/\sqrt{n}} \sim t(n-1) \qquad (\text{assume } H_0\text{: } \mu = \mu_0 \text{ is true})$$

This is again a proper test statistic, which can be used to derive the rejection region; e.g. for a two-tailed test we get:

$$(-\infty, -t_{n-1\,;\,\alpha/2}\,] \cup [\,t_{n-1\,;\,\alpha/2},\ \infty)$$

From this point on, the procedure remains unchanged; if the observed value of the test statistic in a sample falls within the rejection region, the null hypothesis will be rejected (and otherwise not). This test is called the t-test. (Note that the z-test in the previous section still allowed for two test statistics, the sample mean and the standardised sample mean. But here we have no choice, because there is no way we can write the distribution of the sample mean without involving the unknown σ).

The p-value can be found analogously as in the previous section, for example for a right-tailed test:

$$\text{p-value} = P\left(T_{n-1} \geq t_{obs}\right) \quad \text{with } t_{obs} = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}.$$

The exact calculation of p-values is usually only possible by using a computer. Without a computer we can only try to find an interval of possible values for the p-value (using table in the appendix):

<u>*Example 8.18*</u>

Consider the test H_0: $\mu = 500$ versus H_a: $\mu > 500$. In a random sample with 25 observations, the sample mean is found to be 501.2, with a sample standard deviation of 4. At a value of $\alpha = 0.10$, the rejection region for the test statistic above is: [1.318, ∞) (see table 3). The observed value is:

$$\frac{\overline{x}-\mu_0}{s_x/\sqrt{n}} = \frac{501.2-500}{4/\sqrt{25}} = 1.5$$

Because this value falls within the rejection region, the null hypothesis is rejected, and the conclusion will be that there is sufficient proof to state that the population mean is more than 500.

This conclusion could have been found using the p-value: $P(T_{24} \geq 1.5)$.

Table 3 for the t-distribution shows us only that $P(T_{24} \geq 1.318) = 0.10$ and $P(T_{24} \geq 1.711) = 0.05$. Thus, the p-value $P(T_{24} \geq 1.5)$ must be somewhere between 0.05 and 0.10. This tells us immediately that we should not reject H_0 at $\alpha = 0.05$, but we should do so if $\alpha = 0.10$. ◄

Remark. The calculation of the power for t-tests is more difficult than for z-tests. This has to do with the fact that the alternative hypothesis only relates to μ, but that the calculation of the power (and thus also of the probability of a Type II error) involves an assumption about the population variance as well. We will however not dwell on this situation.

8.4.3 The hypothesis σ = σ₀ (normal distribution)

There are cases where we are primarily interested in testing the value of the population standard deviation (or the variance). The steps are performed analogously to the previous hypothesis tests; for a two-sided alternative hypothesis, we will get:

H_0: $\sigma = \sigma_0$ against H_a: $\sigma \neq \sigma_0$ (or, equivalently: H_0: $\sigma^2 = \sigma_0^2$ against H_a: $\sigma^2 \neq \sigma_0^2$)

The relevant test statistic with distribution (see section 7.3.3):

$$\chi^2_{n-1} = \frac{(n-1)\,S^2}{\sigma_0^2} \sim \chi^2(n-1)$$

From $P(\chi^2_{n-1;\,1-\alpha/2} < \dfrac{(n-1)S^2}{\sigma_0^2} < \chi^2_{n-1;\,\alpha/2}) = 1-\alpha$ it follows directly that the rejection region for the above test statistic is:

$$[0, \chi^2_{n-1;1-\alpha/2}] \cup [\chi^2_{n-1;\,\alpha/2}, \infty)$$

Example 8.19

Mr. L. Amp, who loves to have his house well-lit, has a lamp in his living room with 20 halogen bulbs. He notices that he regularly has to take a stepladder to replace a broken bulb. Mr. L. Amp is an econometrist, and considers whether it will cost less effort in the long run if he replaces all the bulbs at the same time on a regular basis, even though they are not broken yet. This will only be a sensible strategy, if all bulbs last approximately equally long, in other words if the standard deviation of the lifespan of the bulbs is small. He calculates that if the standard deviation is less than 500 hours, it is indeed better to choose the new strategy. He designs a test to decide whether he should change his strategy. But he wants to be quite sure of himself, before he will adopt to the new strategy. That's why he chooses as hypotheses

H_0: $\sigma = 500$ versus H_a: $\sigma < 500$

with a level of significance of 5%. He also decides on taking a sample of n=10 bulbs.

The rejection region for this left-tailed test will be $(0, \chi^2_{9,\,0.95}]$. A table for the chi-square distribution shows that the critical value is $\chi^2_{9;\,0.95} = 3.325$.

Only at this stage, mr. L. Amp will observe the sample, in which he carefully measures the lifespans of the 10 bulbs. The sample standard deviation s of those 10 bulbs appears to be 400 hours. Now the observed value for the test statistic can be determined:

$$\frac{(n-1)s^2}{\sigma_0^2} = \frac{9\cdot 400^2}{500^2} = 5.76$$

This outcome does not fall within the critical region, and therefore Mr. L.Amp concludes that there is not sufficient reason to decide to change his strategy based on this sample.

Of course, we could also have used the p-value:

$\text{p-value} = P(\chi_9^2 \le 5.76)$. Using the table for the chi-square distribution, we can only say that the p-value is larger than 0.10 (check; using a computer, we would have found a probability of 0.236). So at $\alpha = 0.05$, the null hypothesis cannot be rejected. ◄

Example 8.20

Consider the test H_0: $\sigma^2 = 7$ versus H_a: $\sigma^2 \neq 7$.
In a sample with 10 observations, the observed sample variance is $s^2 = 1.36$, which results in the observed value for the test statistic of (10–1)*1.36/7 = 1.75.
At this two-tailed test, it is clear that the value of 1.75 is in the left tail of the chi-square distribution, so we get:

$$\text{p-value} = 2P\left(\chi_{n-1}^2 \le \frac{(n-1)\,s^2}{\sigma_0^2}\right)$$

$$= 2P\left(\chi_{10-1}^2 \le \frac{(10-1)\,1.36}{7}\right) \approx 2P(\chi_9^2 \le 1.75)$$

$$= 0.0104 \text{ (using a computer)}$$

So, if $\alpha < 0.0104$, then H_0 cannot be rejected, but whenever $\alpha > 0.0104$ it should be.

If you would have (incorrectly) multiplied the probability $P(\chi_9^2 \ge 1.75)$ by 2, then this would have resulted in a p-value of $2P(\chi_9^2 \ge 1.75) = 1.9896$. This 'probability' is larger than 1, which immediately indicates that something went wrong, and you should have taken $P(\chi_9^2 \le 1.75)$ ◄

8.4.4 The hypothesis p = p₀ *(binomial distribution)*

To test the null hypothesis $p = p_0$, we can choose the number of successes X in the sample. Under H_0, we have $X \sim \text{Bin}(n, p_0)$. This has been shown already in Example 8.8 and Example 8.9. Instead of using this exact distribution, we can also use the normal approximation to the binomial distribution. When n is large enough, this is what is done usually. As rule of thumb, it is often said that $np_0 \ge 5$ and $n(1-p_0) \ge 5$. As shown in section 7.3.5 , we get for $\hat{P} = \frac{X}{n}$ under the null hypothesis $p = p_0$:

$$\hat{P} \overset{approx}{\sim} N\left(p_0, \frac{p_0(1-p_0)}{n}\right)$$

So $Z = \dfrac{\hat{P} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}} \overset{approx}{\sim} N(0,1)$.

So if the alternative hypothesis has the form H_a: $p > p_0$, then the rejection region for the test statistic above will be: $[z_\alpha, \infty)$ (at a level of significance α).

Example 8.21 The triangle test

In an experiment the sugar content of a soft drink is reduced slightly. One wonders whether the modified product can be distinguished from the original one. A random sample of n subjects are all offered three samples of the soft drink in identical bottles. Two samples contain the original product, the third contains the modified one. Each of the n subjects will be asked which of the three samples

contains a different product from the other two. In case they do not notice any difference at all, they are asked to guess. Because the presentation of three test samples to each of the subjects is typical for this type of test, it is called a triangle test.
If the difference is really imperceptible, and everyone simply guesses, then we would expect about a third of the subjects to identify the modified product (even if that happened just by chance). On the other hand, if some subjects are really able to taste the difference, then more than a third of all subjects are expected to identify the modified product, which might indicate a noticeable difference.
We define first:

p = probability that an arbitrary subject will identify correctly the modified product

This leads to the following hypotheses:
H_0: $p = \frac{1}{3}$ versus H_a: $p > \frac{1}{3}$

This is therefore a one-tailed test. Say that the sample consists of 300 people. The test statistic (with distribution) is then:

$$Z = \frac{\hat{P} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{\hat{P} - \frac{1}{3}}{\sqrt{\frac{\frac{1}{3}(1-\frac{1}{3})}{300}}} \sim N(0,\ 1)$$

With associated rejection region (at α=0.05): $[1.645,\ \infty)$.

Now assume that 117 subjects identified the modified product correctly, then the sample proportion is $\hat{p} = 117/300 = 0.39$, and the observed value of Z is $z = 2.08$. That value falls within the rejection region, so there exists sufficient proof for the statement that the modified product can be distinguished from the original!
The p-value would have resulted in: $P(Z \geq 2.08) = 0.0188$. So also at $\alpha = 0.02$, the H_0 should be rejected, but at $\alpha = 0.01$ not anymore. ◄

8.5 Hypotheses concerning the difference between two populations

8.5.1 The hypothesis $\mu_1 = \mu_2$ (normal distributions)

(B&E, pages 403)
We start here by assuming that the standard deviations of the two populations are known. We will base the test statistic on the difference between the two sample means, so $\bar{X}_1 - \bar{X}_2$. We know (under certain requirements, see section 7.4.2):

$$\bar{X}_1 - \bar{X}_2 \sim \mathrm{N}(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}) \quad \text{or} \quad \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \mathrm{N}(0,1)$$

where n_1 and n_2 are the two sample sizes. To test the null hypothesis H_0: $\mu_1 = \mu_2$ (or, equivalently: $\mu_1 - \mu_2 = 0$), we use the following test statistic:

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

which has under the null hypothesis a standard normal distribution.

Example 8.22

In the food industry, machines are used to fill bottles, jars and cans. These machines can be adjusted to certain filling quantities by setting the mean volume (or weight). No two bags of potatoes will weigh exactly the same. The spread which is inevitable is expressed in the variance or the standard deviation. The standard deviation (for a specific product) is therefore a characteristic of the machine used. Say two machines are filling bags with nominally 5 kilograms of potatoes, with $\sigma_1 = 28$ grams and $\sigma_2 = 45$ grams. Now suppose that we take a sample of four bags from both machines, which results in two sample means. Now the standard deviation of the difference of the two sample means is

$$\sigma_{\bar{X}_1-\bar{X}_2} = \sqrt{\frac{28^2}{4}+\frac{45^2}{4}} = 26.5 \quad \text{gram}$$

When we set up a test of the null hypothesis that the mean filling weights of both machines are equal, so H_0: $\mu_1 - \mu_2 = 0$ against a two-sided alternative with $\alpha = 0.05$, we obtain the rejection region for the test statistic $Z: \ (-\infty, \ -1.96] \cup [1.96, \ \infty)$. Now let the two sample means be equal to 4989 and 4951 gram respectively, then the outcome of the test statistic is:

$$\frac{4989-4951}{26.5} = \frac{38}{26.5} = 1.434$$

At this sample, the hypothesis $\mu_1 = \mu_2$ should not be rejected. ◄

The same method can be applied to test H_0: $\mu_1 - \mu_2 = \Delta_0$ (above, we used $\Delta_0 = 0$). So, under this null hypothesis

$$Z = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}}} = \frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}}}$$

has a standard normal distribution. This test can be applied, for example, if one wants to find proof if a certain machine fills bags with at least 100 grams more than another machine does.

If we no longer can assume the standard deviations to be known, but only that they are equal to each other, so $\sigma_1 = \sigma_2$, we can find in section 7.4.3 the derivation of the test statistic:

$$T = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\dfrac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}\left(\dfrac{1}{n_1}+\dfrac{1}{n_2}\right)}} = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{S_P^2\left(\dfrac{1}{n_1}+\dfrac{1}{n_2}\right)}} \sim t_{n_1+n_2-2}$$

For the null hypothesis H_0: $\mu_1 - \mu_2 = \Delta_0$, we obtain:

$$T = \frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{S_P^2\left(\dfrac{1}{n_1}+\dfrac{1}{n_2}\right)}} \sim t_{n_1+n_2-2}$$

This can be used again in the usual way to find rejection regions, p-values etc.

The situation that the standard deviations are not known, and moreover cannot be assumed to be equal to each other, can be developed by the reader (see also § 7.4.4). The problem which arises here is that it is not possible to find a test statistic for which the distribution under H_0 is known. Therefore, a rejection region cannot be determined before the sample is analysed, because the "appropriate" number of degrees of freedom can be determined only after the sample has been analysed. However, the method by determining the p-value can still be used without any problem.

In the next section, a hypothesis test will be discussed which can be used to test the assumptions that both standard deviations are equal to each other, so H_0: $\sigma_1 = \sigma_2$ versus H_a: $\sigma_1 \neq \sigma_2$. That test is regularly used as a first step in evaluating the difference between two population means: its outcome tells us which test should be applied. So, if we reject H_0, then we use a test for the difference between μ_1 and μ_2 without using the assumption that $\sigma_1 = \sigma_2$. But what happens if the null hypothesis cann*ot* be rejected? We have stated earlier that not rejecting H_0 does not at all provide proof that H_0 is true.

Nevertheless, experience has taught statisticians that in this particular case we can say that the differences between σ_1 and σ_2 are probably not so large that a test assuming $\sigma_1 = \sigma_2$ becomes invalid. In other words: if we do not reject H_0: $\sigma_1 = \sigma_2$, then we will simply continue using the assumption of equality.

8.5.2 The hypothesis $\sigma_1 = \sigma_2$ (normal distributions)

(B&E, ex. 8.4.1)
We will consider now the null hypothesis H_0: $\sigma_1^2 = \sigma_2^2$ (or $\sigma_1 = \sigma_2$, or $\sigma_1/\sigma_2 = 1$). In section 7.4.5 we found that:

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{S_1^2\sigma_2^2}{S_2^2\sigma_1^2} \sim F(n_1-1, n_2-1)$$

(so an F-distribution with $n_1 - 1$ (numerator) and $n_2 - 1$ (denominator) degrees of freedom.

By substituting the ratio $\sigma_1 / \sigma_2 = 1$ (see H_0), we obtain the test statistic

$$\frac{S_1^2}{S_2^2} \sim F(n_1-1, n_2-1).$$

The associated rejection region for a two-tailed test with significance level α is then (check!):

$$\left[0\,,\ f_{n_1-1,n_2-1;1-\alpha/2}\right] \cup \left[f_{n_1-1,n_2-1;\alpha/2}\ ,\ \infty\right)$$

See section 7.4.5 for a discussion about the use of F-tables and the swapping of the number of degrees of freedom, such that the rejection region can also be written as:

$$\left[0\,,\ \frac{1}{f_{n_2-1,n_1-1;\alpha/2}}\right] \cup \left[f_{n_1-1,n_2-1;\alpha/2}\ ,\ \infty\right)$$

If we start with the null hypothesis $H_0 : \sigma_1^2/\sigma_2^2 = a_0$ (where a_0 does not have to be 1), the test statistic becomes $\dfrac{S_1^2}{S_2^2 a_0} \sim F(n_1-1, n_2-1)$.

Example 8.23

Test if the standard deviation of a certain population '1' is more than twice the standard deviation of another population '2'. (both populations normally distributed). We test $H_0 : \sigma_1 = 2\sigma_2$ vs $H_1 : \sigma_1 > 2\sigma_2$. Rewritten to variances: $H_0 : \sigma_1^2 = 4\sigma_2^2$ vs $H_1 : \sigma_1^2 > 4\sigma_2^2$, or alternatively: $H_0 : \sigma_1^2/\sigma_2^2 = 4$ vs $H_1 : \sigma_1^2/\sigma_2^2 > 4$.

The test statistic with distribution is:

$$\frac{S_1^2}{4S_2^2} \sim F(n_1-1, n_2-1).$$

If $n_1 = 10$ and $n_2 = 8$, the rejection region (at α = 0.05) will be [3.68, ∞). So with these small samples, the sample variance of the first sample should be more than 14.6 times (= 4*3.68) as large as the sample variance of the second sample, before we are allowed to reject the null hypothesis! ◄

8.5.3 The hypothesis $p_1 = p_2$

Briefly, we will discuss how we can test for the equality of two population proportions of two independent populations. From section 7.4.6 , we know

$$\frac{\hat{P}_1 - \hat{P}_2 - (p_1 - p_2)}{\sqrt{\dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}}} \overset{approx}{\sim} N(0,1)$$

Under the null hypothesis $p_1 = p_2$ ($= p$), we obtain:

$$\frac{\hat{P}_1 - \hat{P}_2}{\sqrt{\dfrac{p(1-p)}{n_1} + \dfrac{p(1-p)}{n_2}}} = \frac{\hat{P}_1 - \hat{P}_2}{\sqrt{p(1-p)}\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \overset{approx}{\sim} N(0,1)$$

This is not a proper test statistic yet, since it still includes an unknown parameter p $(= p_1 = p_2)$. However, this parameter can be estimated by the weighted mean of the two sample proportions (recall that both sample proportions are unbiased estimators for, in this case, the same population parameter):

$$\hat{P}_P = \frac{n_1\hat{P}_1 + n_2\hat{P}_2}{n_1 + n_2} = \frac{X_1 + X_2}{n_1 + n_2}$$

This results in the following test statistic:

$$\frac{\hat{P}_1 - \hat{P}_2}{\sqrt{\hat{P}_P(1-\hat{P}_P)}\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \overset{approx}{\sim} N(0,1)$$

And in case the null hypothesis is stating a difference unequal to 0, so $p_1 - p_2 = \Delta_0$, we can derive the following test statistic:

$$\frac{\hat{P}_1 - \hat{P}_2 - \Delta_0}{\sqrt{\dfrac{\hat{P}_1(1-\hat{P}_1)}{n_1} + \dfrac{\hat{P}_2(1-\hat{P}_2)}{n_2}}} \overset{approx}{\sim} N(0,1)$$

8.6 Problems

8.1 At a fairground booth one sells tickets with the slogan '30% chance of winning a prize!' Someone decides to investigate this claim and buys 50 tickets. He decides that he will confront the vendor about the accuracy of the '30%' claim if he receives less than 10, or more than 20 prizes. Calculate the probability of a type I error.

8.2 The manufacturer of pound bags (500 g) of caster sugar examines the average weight of the pound bags produced by observing a sample of seven bags. He assumes that the fill weight has a standard deviation of 4, and that the weights come from a normal distribution. He decides that he will adjust the production process if the sample will average 498 g or less, or 502 g or more. Calculate the probability of a type I error.

8.3 A method for measuring the concentration of substance G in a liquid has a standard deviation of 1.5 μg/l. One wants to investigate whether the concentration of substance G in a certain sample differs from 45 μg/l. Four measurements were:

45.2 41.3 44.1 42.6

Assume that these measurements can be interpreted as draws from a normal distribution whose unknown expectation μ is the actual concentration.

a Investigate using a test with $\alpha = 0.05$ whether the hypothesis $\mu = 45$ should be rejected.

b For which significance levels will the null hypothesis will rejected?

8.4 One tests $\mu = 500$ with $\alpha = 0.05$. A sample of size 16 is drawn from the $N(\mu, 9^2)$-distribution, and shows a sample mean of 496. Do you reject the hypothesis
a when using a two-sided alternative hypothesis?
b when using a left-sided alternative hypothesis?
c when using a right-sided alternative hypothesis?

8.5 A device measures the concentration of substance A in a liquid (μg/l). The (normally distributed) random error of this device has a standard deviation of 7 μg/l. Ten observations from the same sample are: 165, 188, 183, 181, 178, 169, 194, 181, 176, 175 (compare with exercise 7.7) the 90%-confidence interval for μ was (175.36, 182.64). The 95%-confidence interval was (174.66, 183.34). Will the hypothesis $\mu = 175$ be rejected (using a two-sided test)
a with $\alpha = 0.10$?
b with $\alpha = 0.05$?
c with $\alpha < 0.05$?

8.6 Assume that for a sample of 24 observations from a normal distribution with known standard deviation σ the sample mean $\bar{x}$ equals 3.
a For which value of σ is the 95%-confidence interval for μ equal to (1.0, 5.0)?

Draw in a graph with $\bar{x}$ along the horizontal axis and μ along the vertical axis the lines $\mu = \bar{x}$, $\mu = \bar{x} + 2$ and $\mu = \bar{x} - 2$.

b How can we easily read from this graph the 95%-confidence interval for an arbitrary sample mean?
c How can we easily read from this graph the 95%-prediction interval for the sample mean given an arbitrary μ_0?
d How can we easily read from this graph the rejection region for the hypothesis $\mu = \mu_0$ (with $\alpha = 0.05$)?

8.7 It is investigated whether Lake Pleister is more polluted than before. The concentration of substance C was once 15 mg / m^3. Measurements of concentration have a normal distribution with a standard deviation of 2.6 (mg / m^3). Four measurements are made.
a Determine the critical region for the sample mean with $\alpha = 0.05$.
b Calculate the power of the test when the concentration has increased to 18 μg per m^3.

8.8 Suppose that of a normally distributed random variable the expectation μ is unknown, but $\sigma^2 = 900$ is known. To test H_0: $\mu = 350$ against H_a: $\mu > 350$, one has chosen as a rejection region for the mean of a sample of 16 observations the interval $[365, \infty]$.
a Calculate the probability of an type I error.
b Calculate β if the true value of μ is equal to 372.5.

8.9 Suppose, we have a normal distributed random variable with unknown μ, while $\sigma = ½$ is known. We want to test H_0: $\mu = 8$ against H_a: $\mu \neq 8$ with $\alpha = 0.05$, and take a sample of size 50.
a Determine the critical region for the sample mean, and draw your conclusion when $\bar{x}$ turns out to be 7.8.
b Calculate β if the true value of μ is equal to 7.7.
c Determine the p-value for the mentioned outcome in part a.

8.10 Determine for the following situations whether there is a single or composite null hypothesis.
a Hypothesis: a die is fair.
b Hypothesis: a coin is fair.
c Hypothesis: two experiments have the same probability of success.
d Hypothesis: in each of the *n* city districts the average annual income is equal to 50 k€.
e Hypothesis: in each of the three city districts the average annual income is the same.
f Hypothesis: at intersection A, an average of 15 accidents occurs per month and at intersection B twice as much.
g Hypothesis: at intersections A and B on average the same number of accidents occurs per month.
h In an experiment, three outcomes can occur S, M, and T. Hypothesis: the probability of S equals the probability of M.

8.11 Four measurements of contamination with substance G are:

45.2 41.3 44.1 42.6

- a Use a test with $\alpha = 0.05$ to determine whether the hypothesis $\mu = 45$ should be rejected (with a two-sided alternative); assume that the measurements come from a normal distribution with mean μ.
- b Determine the p-value for the test in part a.
- c Determine the 95%-confidence interval for μ. Does the interval contain the value 45?
- d Given the answer for part c, determine whether the null hypothesis $\mu = 46.5$ should be rejected.
- e Suppose one may assume that the population standard deviation is equal to 1.7068. Calculate again the p-value for the test in part a. Will the null hypothesis (with $\alpha = 0.05$) be rejected? Is this answer different from the one for part a? Explain!

8.12 The production of packages of butter, which must contain an average of 250 grams, is controlled by the manufacturer regularly. On a day, a sample of nine packages gives an average of 243.6 and a standard deviation of 5.710.

- a Use a test with $\alpha = 0.05$ to find a convincing reason to adjust the process. (Assume a normal distribution).
- b For which values of α will the null hypothesis be rejected?

8.13 In a university restaurant, staff and students spent on average € 3.60 per person, with a standard deviation of € 2.00. These expenditures are (approximately) normally distributed. The cafeteria manager wants to increase spending and therefore will run an advertising campaign. Following this advertising campaign, a random sample of 12 expenditures is taken. This sample has an average of € 4.10 and a standard deviation of € 1.20.

- a. Perform a statistical test to assess whether the manager has succeeded in his endeavour, that is, the average spending per person has indeed increased. Take a significance level of 5%.
- b Determine the p-value, and use this value to assess again whether the null hypothesis from part a can be rejected.
- c. Test whether the standard deviation of the expenses has changed ($\alpha = 0.10$). Use the method of rejection regions, then by using the p-value, and finally by determining a $100(1-\alpha)$%-CI.

8.14 In order to function, light bulbs contain a small amount of mercury, about 3-5 mg. A manufacturer of these lamps uses an average of 4 mg per lamp. It may be assumed that the amount of mercury used in a power lamp is normally distributed. For environmental reasons, the producer does not want to use too much mercury. After a maintenance of the production machine, it is suspected that the average amount of mercury in the lamps is currently more than 4 mg. For inspection, 20 lamps are randomly taken and the amount of mercury is measured: this gives an average 4.28 mg with a standard deviation of 0.62 mg.

- a Perform a statistical test with $\alpha = 0.05$ to determine whether the above presumption can be confirmed.
- b Assume that the population standard deviation is known and equal to 0.5 mg. Again, the researchers perform a test where the null hypothesis will be rejected if the sample mean (with $n = 20$) is greater than 4.184 mg. What is the power of the test if $\mu = 4.5$ mg?

8.15 In a survey one has collected observations from a normal distribution:

17, 15, 20, 19, 21, 19, 18, 20, 19, 16.

Perform the following tests with $\alpha = 0.05$.

- a Test the hypothesis H_0: $\mu = 20$ against H_a: $\mu \neq 20$. Perform the test if $\sigma^2 = 2$.
- b Test the hypothesis H_0: $\mu = 20$ against H_a: $\mu \neq 20$. Perform the test if σ^2 is unknown.
- c Test the hypothesis H_0: $\sigma^2 = 2$ against H_a: $\sigma^2 \neq 2$.

8.16 A sample of nine packages of butter gives an average of 243.6 and a standard deviation of 5.710.

- a Determine the p-value for testing $\mu = 250$ against $\mu \neq 250$.
- b Determine the p-value for testing $\sigma = 3$ against $\sigma \neq 3$.

8.17 Suppose an experiment has a probability p of success. The experiment is performed 100 times to test the null hypothesis $p = ½$ against the alternative $p \neq ½$. Use the number of successes X as the test statistic. Determine the rejection region for X and the probability of a type I error if the significance level α equals 0.05.

8.18 If the sample of the triangle test (see Example 8.21) consists of 180 subjects, and 69 of them point out the different product correctly, what would be the conclusion at a significance level of 0.05?

8.19 To see if a random-number generator works reasonably well, in a sample of 100 generated numbers one counts the even numbers. That turns out to be 44. Formulate the hypotheses and calculate the p-value of the observed outcome.

8.20 One wants to know which part of the group (=population) of twelve year old cars gets a broken brake cylinder within half a year after the last APK (General Periodic Inspection for cars in the Netherlands). It used to be 17% before the introduction of the APK. One hopes to show that the percentage is now less. A sample of 640 cars counted 80 with such a defect. Test the hypothesis that the percentage has decreased, with $\alpha = 0.05$. Go through the following steps:

1 Formulate the hypotheses.
2 Describe the assumptions.
3 Determine the test statistic and its distribution.
4 Determine the critical region for this test statistic.
5 Determine the outcome of the test statistic and the conclusion about the decision about the null hypothesis.
6 Formulate the conclusion.
7 Calculate the p-value as well.
8 May we conclude that the APK has a beneficial effect when the null hypothesis is rejected?

8.21 The notion is that a fraction p = 80% of the older owners (> 40 years) of a video recorder does not know how to program it. To investigate this, 25 of such people are questioned.

a As rejection region we could use: [0 , 16], i.e. if the number X (owners who do not know how to program their video recorder) is lower than or equal to 16, then we reject H_0: $p = 0.80$ if favour of the alternative $p < 0.80$. Determine the probability of type I error = $P(X \leq 16 \mid p = 0.80)$.

b1 Suppose that the assumption $p = 0.80$ is *incorrect*, and that the true value of p is equal to 0.7. Determine the power of the test, i.e. $P(X \leq 16 \mid p = 0.70)$.

b2 For each value of p lower than 0.8 (so e.g. if $p = 0.7$) the hypothesis is false. It would be best if in an research this false hypothesis is indeed rejected. However, if the outcome of X is 17 or above, the hypothesis will not be rejected, according to the rejection region mentioned in part a, which results in: a type II error. Calculate the probability of a type II error if p = 0.7.

c Draw β (the probability of a type II error) as function of p on the interval [0, 0.8).

8.22 The notion is that, when a particular drug is used, 2/3 of the patients are cured while only 1/3 is cured when this drug is not used. To prove this, one tests the hypothesis H_0: p = 1/3 against the (single) alternative H_a: p = 2/3, where p is the true but unknown probability of cure in case the drug is used. One decides to take 10 patients, and reject H_0 in favour of H_a ("the drug works") if at least five patients are cured.

a Determine the probability of a type I error, that is $P(X \geq 5)$ if $p = {}^1/_3$. [Consider that $X \geq 5$ leads to 'H_0 is rejected', but $p = {}^1/_3$ means 'H_0 is true', so the conclusion is not in accordance with reality.] Use a table (on Canvas) or a computer, instead of the normal approximation.

b Determine the probability of a type II error, i.e. the probability that H_0 is not rejected (in case $X < 5$), while in reality $p = {}^2/_3$ (and therefore H_0 is false).

c Answer the same questions in case the sample consists of 25 patients and H_0 ($p = {}^1/_3$) is rejected in favour of H_a ($p = {}^2/_3$) if $X \geq 12$.

8.23 [Exam, July 2011] A modern ice-cream vendor is considering to apply for permission from the municipality to place its mobile sales unit on nice summer days at the intersection of some tourist cycle routes. However, he only wants to do that if there are on average more than 250 cyclists passing the intersection per hour. He wants to test this with a significance level of 1%. Assume that the number of cyclists passing per hour is approximately normally distributed with a value of the variance that is three times the value of the population mean. Last summer, during 17 randomly chosen hours, the ice-cream vendor counted the number of cyclists passing through the intersection, finding an average of 267 cyclists per hour.

a Should the ice-cream vendor apply for the permit? Give (i) test statistic with distribution, (ii) rejection region, (iii) observed value, (iv) decision (yes/no H_0 reject) and (v) conclusion (in words).

b Determine the p-value that belongs to this test and sample.

c Determine the power of this test if the true expected number of cyclist is 260.

d Give two ways in which the ice-cream vendor could have increased the power of the test (from part c).

8.24 Someone doubts whether a die is fair, but is only interested in whether the chance of throwing a five is equal to the chance of a six. He takes this as null hypothesis. He throws 200 times, finds 27 fives and 40 sixes.

a Show that the number of fives X, assuming that the total number of fives and sixes $X + Y = m$, has a BINOM(m, ½)-distribution, in case the null hypothesis is true.

b Determine the p-value. (Use the normal approximation.)

8.25 Two filling processes follow a normal distribution. The standard deviation of the filling amount is equal to 8 ml for machines A and B. The means should be around 4 litres. A sample is available from both machines:

A	3997 4007 4018 4006 4005 4001 4008
B	4006 4018 3994 3999 4000 4001

Test (with $\alpha = 0.05$) the hypothesis that the mean filling amounts for both machines are equal to each other.

8.26 Cat food cans are filled by two machines X and Y, whose standard deviations are 6 and 8 grams respectively. Assume that X and Y are normally distributed. Two samples are given:

X	326 309 317 313 317 321 330 319
Y	324 337 315 324 319 328

Look for a possible difference in filling volume. Give the p-value.

8.27 Of two samples of 13 observations from normal distributions, the averages are 1005 and 1001. Furthermore, it is assumed that for both distributions: $\sigma = 4$.

a Test the hypothesis that the population means are equal (two-sided with $\alpha = 0.05$).

b Determine the power of the test in part a if the true difference of the population means $\mu_1 - \mu_2$ is equal to 3. (Hint, first determine the rejection region for the difference of the two population means).

c Determine for both means separately a 95%-confidence interval (see also exercise 7.26).

d Note that both confidence intervals overlap each other, but the hypothesis $\mu_1 = \mu_2$ is still rejected. How is this possible?

e Test the hypothesis $\mu_1 - \mu_2 = 2$, two-sided with $\alpha = 0.05$.

8.28 Two samples of n observations are known to have a difference in means of 2.3. The standard deviation of both distributions is equal to 9. How big should n be at least to reject the hypothesis (with a = 0.05) that both expectations are equal?

8.29 Two samples from normal distributions with the same variance are as follows:

Sample I	27.1	23.6	25.0	26.2	25.2	
Sample II	28.7	24.8	25.5	26.9	27.2	25.9

Test the hypothesis that the means of both distributions are equal. Use a significance level of 0.01. Formulate the conclusion.

8.30 One examines the influence of environmental pollution on the growth of young conifers. To this end, large plant bins are all provided with the same soil mix. Two bins, each with six trees, are placed in an area where a lot of SO_2 is present in the air, and two similar bins in a clean air area. After five years, the increase in the thickness of the strains is measured. These increases may be considered as independent, normally distributed variables with the same variance. The results of the measurements are (in mm):

Area with clean air						Area with a lot of SO_2					
37	38	32	36	29	38	35	33	32	37	34	29
38	34	41	33	36	34	30	28	31	30	33	36

Investigate whether the amount of SO_2 influences the tree growth, as follows:

a Formulate the null hypothesis and the alternative hypothesis.

b Give test statistic, the distribution and critical region for $\alpha = 0.05$.

c Determine the outcome of the test statistic and draw your conclusion.

8.31 Two machines fill bottles with olive oil. The volumes are assumed to have normal distributions with equal variances. One wants to know if there is a difference between the expected volumes on both machines. The results of two samples (one sample for each machine) are (volumes in ml):

I	326	309	317	313	317	321	330	319
II	324	337	315	324	319	328		

Test the hypothesis that the mean volumes are equal. Use $\alpha = 0.05$, and indicate clearly whether you perform a one-sided or two-sided test, and why.

8.32 One wants to know what impact the skewed position of an old-fashioned hourglass has on the time it takes for the sand to pass to the lower half. To this end, the hourglass is placed in two positions: respectively at an angle of 10 and 20 degrees with the vertical axis. In both positions six measurements have been made. Assume equal variances and normal distributions. The measured times are (in seconds):

Angle 10°	184.4	186.1	187.5	188.1	181.1	185.8
Angle 20°	186.1	181.7	180.5	181.9	180.8	177.1

Use a statistical test to check if there is a difference in the expected throughput times in the two different positions. Use 0.05 as significance level.

8.33 Ten patients are subjected to two treatments X and Y. One finds:

Patient	1	2	3	4	5	6	7	8	9	10
X	47	38	50	33	47	23	40	42	15	36
Y	52	35	52	35	46	27	45	41	17	41
Difference (D)	−5	+3	−2	−2	+1	−4	−5	+1	−2	−5

Test H_0: $\mu_X = \mu_Y$ against H_a: $\mu_X \neq \mu_Y$ with $\alpha = 0.10$, assuming that the variances of the treatments are equal. (What assumptions are required to perform the test?) What is your conclusion in words?

8.34 To investigate whether there is a difference between two measurement methods, five samples of a certain (homogeneous) substance are each divided into two parts. On one half method I is applied, on the other half method II, with the result:

Method \ Sample:	1	2	3	4	5
I	11.0	2.0	8.3	3.1	2.4
II	11.2	1.9	8.5	3.3	2.4
Difference	−.2	.1	−.2	−.2	0

a Justify the choice between a one-sided or two-sided test.
b Test – under the assumption of normality of the distribution(a) – with $\alpha = 0.05$ and based on the given samples, whether the measurement methods differ systematically.

8.35 Two tennis players have a discussion about the weight of the balls before and after use. One keeps arguing that due to hair loss during a tennis match the balls lose weight, while the other believes that this effect is compensated, or even surpassed by the fact that the outside of the ball takes up gravel and dust. Two cans of six brand new balls will be weighed. Of other players, present at the same time at the tennis court, they confiscate used balls of the same brand, and also determine the weight. The weights are:

New:	56.1	56.4	56.3	56.3	56.7	56.5		
Old:	56.2	56.7	56.2	56.1	56.2	56.0	56.3	56.0

Consider these numbers as observations from normal distributions with the same variance and formulate precisely the conclusion based on this observations. Use 0.05 as significance level.

8.36 Draw samples with $n_x = 2n$ and $n_y = n$ from normal distributions with $\sigma_x^2 = \sigma_y^2 = 100$.
Determine n such that, when testing H_0: $\mu_x = \mu_y$ against H_a: $\mu_x = \mu_y + 3$ it holds that $\alpha = \beta = 0.10$.
Naturally, n should be integer. What are the consequences of this?

8.37 To investigate whether the electrical resistance of a wire of a certain material can be reduced by at least 0.05 by applying an alloy, samples of size 12 of both types of wire (alloy and standard) are measured with the following results:

	average	standard deviation
Alloy (L)	0.083	0.003
Standard (S)	0.136	0.002

Test H_0: $\sigma_L^2 = \sigma_S^2$ two-sided with $\alpha = 0.10$. What is your conclusion in words?

8.38 Two machines are acquired to fill vinegar bottles. With both machines, a sample 10 bottles are filled with nominally 1 litre. Using both samples one can determine the systematic deviations from the nominal filling amount of the machines, but in this case the purpose is to compare the variances. The sample standard deviations were for the machines $s_1 = 0.123$ and $s_2 = 0.208$ (ml) respectively.
a Test the hypothesis that the machines are equally accurate. Take 0.05 as significance level.
b Determine 95%-confidence intervals for σ_1 and σ_2 separately. (see also exercise 7.32)
c Determine a 95%-confidence interval for the ratio σ_1 / σ_2 .
d (Multiple choice) The conclusion based on the test in part a can also be deduced from
1 the answers for part b, but not from the answers for part c.
2 the answers for part c, but not from the answers for part b.
3 the answers for part b, as well as from the answers for part c.
4 only by combining the answers for part b and c.

e Suppose that with the first machine *n* bottles are filled and with the second machine 10 bottles are filled. For which minimum value of *n* is the hypothesis from part a rejected in case the sampling standard deviations are the same as above?

f Again, determine the 95%-confidence interval for σ_1 / σ_2, but now with the sample size found at part e as the sample size of the first sample.

8.39 Two samples from normal distributions should shed light on any differences in variance of the underlying populations. One wants to investigate H_0: $\sigma_X^2 = \sigma_Y^2$. The descriptive statistics of the samples are:

	Sample size	sample mean	sample variance
X	12	235	421
Y	9	286	511

a Test H_0: $\sigma_X^2 = \sigma_Y^2$ two-sided with $\alpha = 0.10$. What is your conclusion in words?

b Determine a 90%-confidence interval for σ_X^2 / σ_Y^2. What is the connection with the conclusion in part a?

c Test H_0: $\mu_X = \mu_Y$ against H_a: $\mu_X < \mu_Y$ with $\alpha = 0.01$. What is your conclusion in words?

8.40 In exercise 7.31 two confidence intervals have been derived, depending on whether or not the standard deviations are equal. Therefore, we would first like to test H_0: $\sigma_X^2 = \sigma_Y^2$. Given the data above, determine the smallest significance level for which the hypothesis of equal variations is rejected.

8.41 One wants to investigate whether the percentage of people who regularly watch a local TV channel is different in cities as compared to the countryside. In two samples, 41% is found as a percentage of 2,000 urban residents and 45% in 2,500 rural people.

a For which significance levels do you conclude that there is no difference?

b For which significance levels do you conclude that the percentage of rural people watching the local TV stations is at least 2 percent points higher than the percentage of urban residents?

Appendix A

A.1 Partial integration

$\int f(x)\, dg(x) = f(x)\, g(x) - \int g(x)\, df(x)$ where: $dg(x) = g'(x)dx$ and $df(x) = f'(x)dx$

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:

$\int u(x)\, v(x)\, dx = u(x)\, V(x) - \int V(x)\, u'(x)\, dx$ where: $\int v(x)\, dx = V(x)$

A.2 Series

$$\sum_{i=1}^{n} i = \frac{n(n+1)}{2}$$

$$\sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}$$

$$\sum_{i=0}^{n} r^i = \frac{1-r^{n+1}}{1-r}$$

$$\sum_{i=0}^{\infty} r^i = \frac{1}{1-r} \quad \text{when } |r|<1$$

$$\sum_{i=1}^{\infty} i r^i = \frac{r}{(1-r)^2} \quad \text{when } |r|<1$$

$$\sum_{i=0}^{\infty} \frac{x^i}{i!} = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + ... = e^x$$

A.3 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: $\Gamma(n) = (n-1)\,\Gamma(n-1) = (n-1)\,(n-2)\,\Gamma(n-2) = (n-1)\,(n-2)\,...\,\Gamma(1) = (n-1)!$

Also (without proof): $\Gamma(\tfrac{1}{2}) = \sqrt{\pi} \;\Rightarrow\; \Gamma(1\tfrac{1}{2}) = \tfrac{1}{2}\sqrt{\pi} \;\Rightarrow\; \Gamma(2\tfrac{1}{2}) = \tfrac{3}{4}\sqrt{\pi}$.

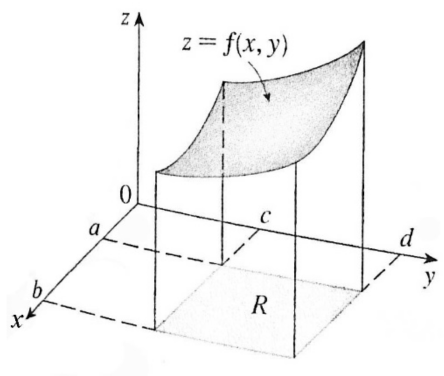
A.4 Greek alphabet

Α α alfa	Η η èta	Ν ν nu	Τ τ tau
Β β bèta	Θ θ thèta	Ξ ξ xi	Υ υ ypsilon
Γ γ gamma	Ι ι iota	Ο ο omikron	Φ φ phi
Δ δ delta	Κ κ kappa	Π π pi	Χ χ chi
Ε ε epsilon	Λ λ lambda	Ρ ρ rho	Ψ ψ psi
Ζ ζ zèta	Μ μ mu	Σ σ sigma	Ω ω omega

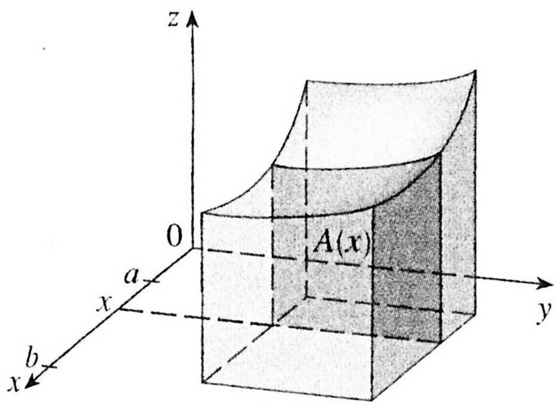
A.5 Multiple integrals

As is well known, the area under the curve $f(x)$ and the horizontal axis from a to b can be represented by the integral $\int_a^b f(x)dx$.

Now we will focus on the volume under the curve $f(x,y)$, where $a \le x \le b$ and $c \le y \le d$.
(see figure). We say we integrate over the rectangle $R = \{(x,y) \mid a \le x \le b \text{ and } c \le y \le d\}$.
(Credits: the figures below were copied from the textbook for the course Mathematics 1: Stewart, J., *Calculus: Early trancendentals).*



Assume now a value of x fixed somewhere between a and b. Then we can integrate $f(x,y)$ with respect to y; the result will be a function of x, and we can write: $A(x) = \int_c^d f(x,y)dy$.

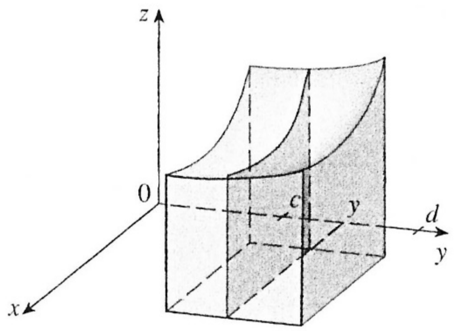


$A(x)$ represents now the area of the intersection of the 3-dimension form under the curve $f(x,y)$ with the vertical plane with a fixed x-value. If we next integrate $A(x)$ with respect to x we get the volume:

$$\text{Volume} = \int_a^b A(x)dx = \int_a^b \left[\int_c^d f(x,y)dy \right] dx = \int_a^b \int_c^d f(x,y)dy\,dx \,.$$

Note that during the evaluation of double integrals, we work from the inside out.

We also could have swapped the order of integration (pay attention to the limits):



$$\text{Volume} = \int_c^d \left[\int_a^b f(x,y)dx \right] dy = \int_c^d \int_a^b f(x,y)dx\,dy$$

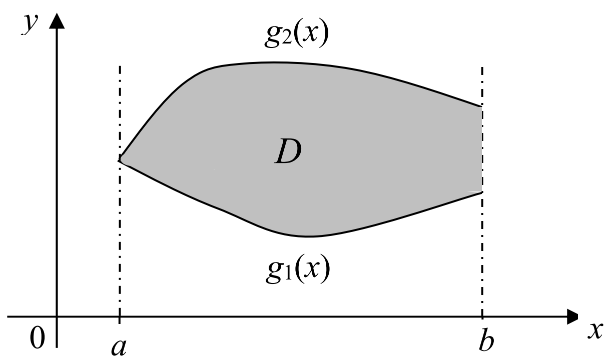
We conclude that $\int_a^b \int_c^d f(x,y)dy\,dx = \int_c^d \int_a^b f(x,y)dx\,dy$.

Remark. This last result is by the way not *always* true; Fubini's Theorem states that the order of integration can be reversed, if the integral of the absolute value of $f(x,y)$ is finite. Fortunately, that will always be the case when we use double integrals to determine probabilities. When determining moments, there might theoretically be a problem; we will however not dwell on these cases.

If $f(x,y)$ can be written as the product $g(x)h(y)$ (we say that $f(x,y)$ has been factorised into a function of x and a function of y), then it follows simply that:

$$\int_a^b \int_c^d g(x)h(y)dy\,dx = \int_a^b g(x) \left[\int_c^d h(y)dy \right] dx = \int_a^b g(x)\,dx \int_c^d h(y)dy$$

When the area of integration is not rectangular (so no Cartesian product), we have to pay extra



attention to the integration limits. Say for example that we need to integrate over the area D in the figure above, and if we, as before, first treat x as a constant and start by integrating with respect to y, then the area of the intersection $A(x)$ is expressed by $A(x) = \int_{g_1(x)}^{g_2(x)} f(x,y)dy$. Thus, we get:

$$\iint_D f(x,y)\,dy\,dx = \int_a^b \int_{g_1(x)}^{g_2(x)} f(x,y)\,dy\,dx\,.$$

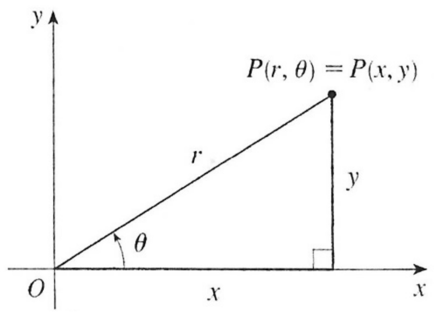
Just as with single integrals, we can substitute variables in double integrals. Consider the 2-dimensional function $f(x,y)$ which has to be integrated over the region *D*. For some functions, this integral becomes much more tractable by working in another system of coordinates. Say we will substitute *x* by $g(u,v)$, and *y* by $h(u,v)$, where the functions *g* and *h* define a 1-to-1 transformation of the (u,v)-plane to the (x,y)-plane. We write *S* for the region in the (u,v)-plane which is transformed to *D*. Then it can be shown (under quite general conditions of continuity) that:

$$\iint_D f(x,y)\,dy\,dx = \iint_S f(g(u,v),h(u,v))\,|J|\,du\,dv$$

where *J* denotes the Jacobian, the determinant of the matrix with all first-order partial derivatives:

$$J = \begin{vmatrix} \dfrac{\partial x}{\partial u} & \dfrac{\partial x}{\partial v} \\ \dfrac{\partial y}{\partial u} & \dfrac{\partial y}{\partial v} \end{vmatrix} = \begin{vmatrix} \dfrac{\partial g(u,v)}{\partial u} & \dfrac{\partial g(u,v)}{\partial v} \\ \dfrac{\partial h(u,v)}{\partial u} & \dfrac{\partial h(u,v)}{\partial v} \end{vmatrix}$$

A very common transformation is the transformation to polar coordinates. Each point in the Euclidian (x,y)-plane can be written by two polar coordinates (r, θ), where *r* is the distance from the origin to the point (x,y), and θ is the angle of the line origin - (x,y) with the positive *x*-axis. Thus, we can write:
$x = r\cos\theta$ and $y = r\sin\theta$. The Jacobian will be equal to *r* (check!).



By way of example, we will use polar coordinates here to prove that the area under the pdf of a standard normal distribution is equal to 1. (In the course Prob. Theory and Statistic 1, we only saw a proof which was based on the fact that $\Gamma(½) = \sqrt{\pi}$, but for this latter equality, no proof was given.)

We want to prove that $B = 1$, where $B = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}\,dx$.

$$\Rightarrow \quad B^2 = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}\,dx \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2}\,dy = \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{1}{2}(x^2+y^2)}\,dy\,dx\,.$$

When we perform the substitution $x = r\cos\theta$ and $y = r\sin\theta$, we have to determine which region *S* in the (r, θ)-plane is transformed to the complete, unrestricted (x, y)-plane (since both *x* and *y* are between $-\infty$ and ∞. This is clearly $S = \{(r,\theta) \mid r > 0 \text{ en } 0 \le \theta \le 2\pi\}$.

Therefore:

$$\Rightarrow \quad B^2 = \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{1}{2}(x^2+y^2)}\,dy\,dx = \frac{1}{2\pi}\int_0^{\infty}\int_0^{2\pi} e^{-\frac{1}{2}(r^2\sin^2\theta + r^2\cos^2\theta)}\,r\,d\theta\,dr$$

$$= \frac{1}{2\pi}\int_0^{\infty}\int_0^{2\pi} e^{-\frac{1}{2}r^2}\,r\,d\theta\,dr = \int_0^{\infty} e^{-\frac{1}{2}r^2}\,r\,dr = -e^{-\frac{1}{2}r^2}\Big|_0^{\infty} = 1$$

which completes the proof.

Appendix B

Table 1. Binomial Distribution

Tabled is: $P(X \le k) = \sum_{x=0}^{k} p(x)$.

n = 5

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

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

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

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**

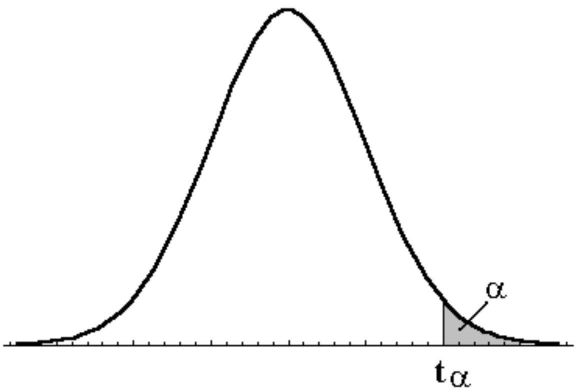
	p														
k	**.01**	**.05**	**.10**	**.20**	**.25**	**.30**	**.40**	**.50**	**.60**	**.70**	**.75**	**.80**	**.90**	**.95**	**.99**
0	.778	.277	.072	.004	.001	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
1	.974	.642	.271	.027	.007	.002	.000	.000	.000	.000	.000	.000	.000	.000	.000
2	.998	.873	.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: $P(X \le k) = \sum_{x=0}^{k} p(x).$

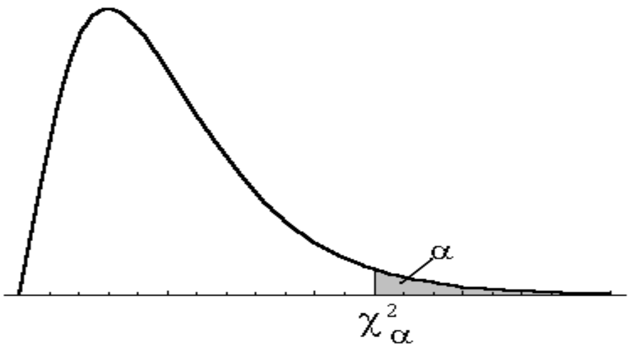
k	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0	6.0	7.0	8.0	9.0	10.0	11.0	12.0	13.0	14.0	15.0
0	.368	.223	.135	.082	.050	.030	.018	.011	.007	.002	.001	.000	.000	.000	.000	.000	.000	.000	.000
1	.736	.558	.406	.287	.199	.136	.092	.061	.040	.017	.007	.003	.001	.000	.000	.000	.000	.000	.000
2	.920	.809	.677	.544	.423	.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. Right-tail critical values for *t*-distribution



Degrees of freedom	$t_{.10}$	$t_{.05}$	$t_{.025}$	$t_{.01}$	$t_{.005}$	Degrees of freedom	$t_{.10}$	$t_{.05}$	$t_{.025}$	$t_{.01}$	$t_{.005}$
1	3.078	6.314	12.706	31.821	63.657	**24**	1.318	1.711	2.064	2.492	2.797
2	1.886	2.920	4.303	6.965	9.925	**25**	1.316	1.708	2.060	2.485	2.787
3	1.638	2.353	3.182	4.541	5.841	**26**	1.315	1.706	2.056	2.479	2.779
4	1.533	2.132	2.776	3.747	4.604	**27**	1.314	1.703	2.052	2.473	2.771
5	1.476	2.015	2.571	3.365	4.032	**28**	1.313	1.701	2.048	2.467	2.763
6	1.440	1.943	2.447	3.143	3.707	**29**	1.311	1.699	2.045	2.462	2.756
7	1.415	1.895	2.365	2.998	3.499	**30**	1.310	1.697	2.042	2.457	2.750
8	1.397	1.860	2.306	2.896	3.355	**35**	1.306	1.690	2.030	2.438	2.724
9	1.383	1.833	2.262	2.821	3.250	**40**	1.303	1.684	2.021	2.423	2.704
10	1.372	1.812	2.228	2.764	3.169	**45**	1.301	1.679	2.014	2.412	2.690
11	1.363	1.796	2.201	2.718	3.106	**50**	1.299	1.676	2.009	2.403	2.678
12	1.356	1.782	2.179	2.681	3.055	**60**	1.296	1.671	2.000	2.390	2.660
13	1.350	1.771	2.160	2.650	3.012	**70**	1.294	1.667	1.994	2.381	2.648
14	1.345	1.761	2.145	2.624	2.977	**80**	1.292	1.664	1.990	2.374	2.639
15	1.341	1.753	2.131	2.602	2.947	**90**	1.291	1.662	1.987	2.368	2.632
16	1.337	1.746	2.120	2.583	2.921	**100**	1.290	1.660	1.984	2.364	2.626
17	1.333	1.740	2.110	2.567	2.898	**120**	1.289	1.658	1.980	2.358	2.617
18	1.330	1.734	2.101	2.552	2.878	**140**	1.288	1.656	1.977	2.353	2.611
19	1.328	1.729	2.093	2.539	2.861	**160**	1.287	1.654	1.975	2.350	2.607
20	1.325	1.725	2.086	2.528	2.845	**180**	1.286	1.653	1.973	2.347	2.603
21	1.323	1.721	2.080	2.518	2.831	**200**	1.286	1.653	1.972	2.345	2.601
22	1.321	1.717	2.074	2.508	2.819	∞	1.282	1.645	1.960	2.326	2.576
23	1.319	1.714	2.069	2.500	2.807						

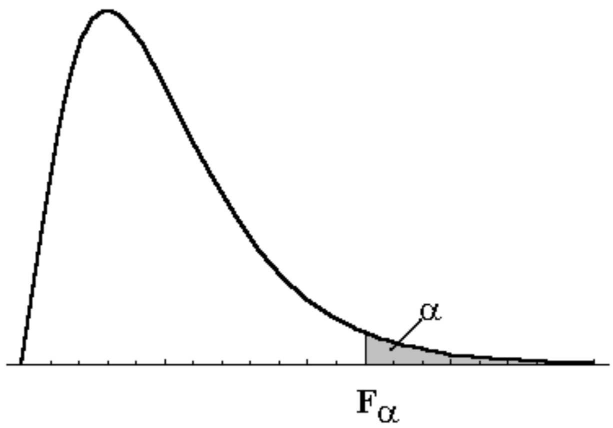
Table 4. Right-tail critical values for chi-square distribution



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 5. Right-tail critical values for *F*-distribution with α =.05

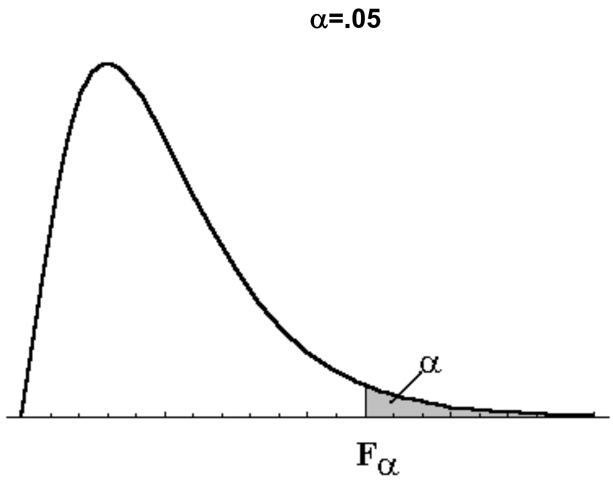
α=.05



NUMERATOR DEGREES OF FREEDOM

DENOMINATOR DEGREES OF FREEDOM df_{den} \ df_{num}	1	2	3	4	5	6	7	8	9
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88

Tabel 5b. Right-tail critical values for *F*-distribution with α =.05



df_{num}	NUMERATOR DEGREES OF REEDOM								
	10	**12**	**15**	**20**	**24**	**30**	**40**	**60**	**120**
df_{den} **1**	241.9	243.9	246.0	248.0	249.1	250.1	251.1	252.2	253.3
2	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49
3	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55
4	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66
5	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40
6	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70
7	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27
8	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97
9	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75
10	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58
11	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45
12	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34
13	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25
14	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18
15	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11
16	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06
17	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01
18	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97
19	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93
20	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90
21	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87
22	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84
23	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81
24	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79
25	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77
26	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75
27	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73
28	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71
29	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70
30	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68
40	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58
60	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47
120	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35
∞	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22

DENOMINATOR DEGREES OF FREEDOM

Table 6. 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

Index

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	$\{0, 1, 2, 3, \ldots, n\}$	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	$\{r, r+1, r+2, \ \ldots\}$	$\frac{r}{p}$	$\frac{r(1-p)}{p^2}$	$\left(\frac{pe^t}{1-qe^t}\right)^r$
Geometric(p)	$p(1-p)^{x-1}$	p	$\{1, 2, 3, \ldots\}$	$\frac{1}{p}$	$\frac{1-p}{p^2}$	$\frac{pe^t}{1-qe^t}$
Poisson(λ)	$e^{-\lambda} \frac{\lambda^x}{x!}$	λ	$\{0, 1, 2, 3, \ldots\}$	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)	$(-\infty\,, \infty)$	μ	σ^2	$e^{\mu t+\sigma^2 t^2/2}$
Exponential(λ)	$\lambda e^{-\lambda x}$	$\lambda > 0$	$[0\,, \infty)$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	$\frac{\lambda}{\lambda-t}$
Gamma(θ, r)	$\frac{x^{r-1}e^{-x/\theta}}{\theta^r\Gamma(r)}$	θ , $r > 0$	$[0\,, \infty)$	$r\theta$	$r\theta^2$	$\left(\frac{1}{1-\theta t}\right)^r$
Chi-square(ν)	$\frac{x^{\nu/2-1}e^{-x/2}}{2^{\nu/2}\Gamma(\nu/2)}$	$\nu = 1, 2, \ldots$	$[0\,, \infty)$	ν	2ν	$(1-2t)^{-\nu/2}$

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	$\{0, 1, 2, 3, \ldots, n\}$	np	$np(1-p)$	$(pe^t + q)^n$
Negative-binomial(r , p)	$\binom{x-1}{r-1} p^r (1-p)^{x-r}$	p , r	$\{r, r+1, r+2, \ldots\}$	$\frac{r}{p}$	$\frac{r(1-p)}{p^2}$	$\left(\frac{pe^t}{1-qe^t}\right)^r$
Geometric(p)	$p(1-p)^{x-1}$	p	$\{1, 2, 3, \ldots\}$	$\frac{1}{p}$	$\frac{1-p}{p^2}$	$\frac{pe^t}{1-qe^t}$
Poisson(λ)	$e^{-\lambda} \frac{\lambda^x}{x!}$	λ	$\{0, 1, 2, 3, \ldots\}$	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)	$(-\infty , \infty)$	μ	σ^2	$e^{\mu t+\sigma^2 t^2/2}$
Exponential(λ)	$\lambda e^{-\lambda x}$	$\lambda > 0$	$[0 , \infty)$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	$\frac{\lambda}{\lambda-t}$
Gamma(θ, r)	$\frac{x^{r-1}e^{-x/\theta}}{\theta^r\Gamma(r)}$	θ , $r > 0$	$[0 , \infty)$	$r\theta$	$r\theta^2$	$\left(\frac{1}{1-\theta t}\right)^r$
Chi-square(ν)	$\frac{x^{\nu/2-1}e^{-x/2}}{2^{\nu/2}\Gamma(\nu/2)}$	$\nu = 1, 2, \ldots$	$[0 , \infty)$	ν	2ν	$(1-2t)^{-\nu/2}$

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	$\{0, 1, 2, 3, \ldots, n\}$	np	$np(1-p)$	$(pe^t + q)^n$
Negative-binomial(r , p)	$\binom{x-1}{r-1} p^r (1-p)^{x-r}$	p , r	$\{r, r+1, r+2, \ldots\}$	$\frac{r}{p}$	$\frac{r(1-p)}{p^2}$	$\left(\frac{pe^t}{1-qe^t}\right)^r$
Geometric(p)	$p(1-p)^{x-1}$	p	$\{1, 2, 3, \ldots\}$	$\frac{1}{p}$	$\frac{1-p}{p^2}$	$\frac{pe^t}{1-qe^t}$
Poisson(λ)	$e^{-\lambda}\frac{\lambda^x}{x!}$	λ	$\{0, 1, 2, 3, \ldots\}$	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)	$(-\infty , \infty)$	μ	σ^2	$e^{\mu t+\sigma^2 t^2/2}$
Exponential(λ)	$\lambda e^{-\lambda x}$	$\lambda > 0$	$[0 , \infty)$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	$\frac{\lambda}{\lambda-t}$
Gamma(θ, r)	$\frac{x^{r-1}e^{-x/\theta}}{\theta^r\Gamma(r)}$	θ , $r > 0$	$[0 , \infty)$	$r\theta$	$r\theta^2$	$\left(\frac{1}{1-\theta t}\right)^r$
Chi-square(ν)	$\frac{x^{\nu/2-1}e^{-x/2}}{2^{\nu/2}\Gamma(\nu/2)}$	$\nu = 1, 2, \ldots$	$[0 , \infty)$	ν	2ν	$(1-2t)^{-\nu/2}$

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	$\{0, 1, 2, 3, \ldots, n\}$	np	$np(1-p)$	$(pe^t + q)^n$
Negative-binomial(r , p)	$\binom{x-1}{r-1} p^r (1-p)^{x-r}$	p , r	$\{r, r+1, r+2, \ldots\}$	$\frac{r}{p}$	$\frac{r(1-p)}{p^2}$	$\left(\frac{pe^t}{1-qe^t}\right)^r$
Geometric(p)	$p(1-p)^{x-1}$	p	$\{1, 2, 3, \ldots\}$	$\frac{1}{p}$	$\frac{1-p}{p^2}$	$\frac{pe^t}{1-qe^t}$
Poisson(λ)	$e^{-\lambda}\frac{\lambda^x}{x!}$	λ	$\{0, 1, 2, 3, \ldots\}$	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	$\max(0, M+n-N) \leq$ $\leq x \leq \min(n, M)$	$\frac{nM}{N}$	$n\frac{M}{N}\frac{N-M}{N}\frac{N-n}{N-1}$	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)	$(-\infty , \infty)$	μ	σ^2	$e^{\mu t+\sigma^2 t^2/2}$
Exponential(λ)	$\lambda e^{-\lambda x}$	$\lambda > 0$	$[0 , \infty)$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	$\frac{\lambda}{\lambda-t}$
Gamma(θ, r)	$\frac{x^{r-1}e^{-x/\theta}}{\theta^r\Gamma(r)}$	θ , $r > 0$	$[0 , \infty)$	$r\theta$	$r\theta^2$	$\left(\frac{1}{1-\theta t}\right)^r$
Chi-square(ν)	$\frac{x^{\nu/2-1}e^{-x/2}}{2^{\nu/2}\Gamma(\nu/2)}$	$v = 1, 2, \ldots$	$[0 , \infty)$	ν	2ν	$(1-2t)^{-\nu/2}$