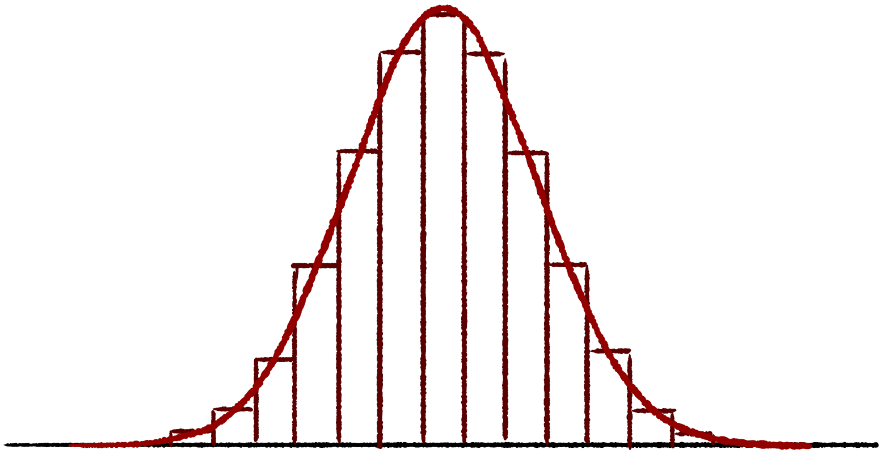
Probability Theory

Version 0.9
April 1, 2020



Mike Vandal Auerbach
www.mathematicus.dk

Probability Theory
Version 0.9, 2020

These notes are a translation of the Danish "Sandsynlighedsregning" written for the Danish stx.

The notes mostly cover only probability theory, although topics cover statistical uses (the sections on the binomial test and samples) because these are most easily explained using probability theory.

The document is written using the document preparation system LaTeX, see www.tug.org and www.miktex.org. Figures and diagrams are produced using *pgf/TikZ*, see www.ctan.org/pkg/pgf.

This document and other resources are available for download from www.mathematicus.dk

Contents

1 Combinatorics

Probability theory is a branch of mathematics which deals with fixing numbers to random phenomena. These might be

- the roll of a dice,
- the winnings on a scratch ticket, or
- the height of a person chosen at random.

The *probabilities* are a measure of how often a phenomenon occurs, e.g. how often we get 5 when we roll a dice.

1.1 In how many ways . . .

If we want to calculate probabilities, it is useful to have formulas which allow us to answer the question "in how many ways is . . . possible?" We can always write down every possibility and the count them all, but this is usually quite cumbersome – especially if there are many ways in which the given outcome might occur.

We therefore start with the following example:

Example 1.1 In a certain restaurant, we can choose between 3 starters, 4 main courses, and 2 desserts. If a menu consists of a starter, a main course, and a dessert, how many menus can we put together?

If we draw a tree diagram of our possible choices, we get figure 1.1.

When we count the lower branches of the tree, we find that there are 24 different possible menus.

There is nothing wrong with finding the answer by drawing trees and counting, but the task quite quickly becomes a mountain of work if there are more than a few choices. We might instead reason like this: For each

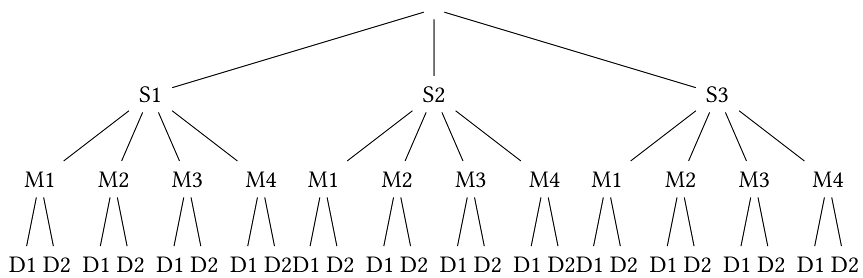


Figure 1.1: A tree diagram of possible menus. "S1" is starter no. 1, $M1$ is main course no. 1, etc.

of the 3 starters, there are 4 main courses to choose from, and for each of those, there are 2 possible desserts. This gives us

$$3 \cdot 4 \cdot 2 = 24$$

different menus.

The calculation in the example shows what we call the *multiplication principle*. Here, the number of possibilities of each *partial choice* (starter, main course, dessert) are multiplied. The keyword here is "and", because we use the principle when we choose a starter *and* a main course *and* a dessert. So, we have:

Theorem 1.2: Multiplication principle

If M contains m elements, and N contains n elements, then we can choose an element of M *and* an element of N in

$$m \cdot n$$

different ways.

There is another principle called the *addition principle* – because we add. We use this principle in situations like these:

Example 1.3 A poor student enters the restaurant from our previous example. He can only afford one dish, so he has to choose between either a starter, a main course, or a dessert. In this case, there are

$$3 + 4 + 2 = 9$$

possibilities to choose a dish because there are 9 dishes in total, and he can only choose one.

The keywords with this principle are "either ... or" because he has to choose *either* a starter, *or* a main course, *or* a dessert.

So, we have

Theorem 1.4: Addition principle

If M contains m elements, and N contains n elements, we can choose an element of M *or* an element of N in

$$m + n$$

different ways.

1.2 Variations

Another question of interest is how to choose a number of elements from a larger set. E.g. if we want to choose 3 elements from a set of 5, in how many ways can we do this?

To set up a general formula, we will need a bit of notation which allows us to write the formulas in a simpler way.

Definition 1.5

If n is a natural number, we define $n!$,

$$n! = n \cdot (n-1) \cdot (n-2) \cdots 2 \cdot 1 \; .$$

$0!$ is defined to be $0! = 1$.
The number $n!$ is called "n factorial".

Example 1.6 The number $6!$ is the number

$$6! = 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 720 \; .$$

As this calculation shows, $n!$ quite quickly becomes a large number, even for small values of n.

To answer the question above, we analyse the following examples:

Example 1.7 At a film night, 5 different films are in play, but there is only time to watch 3 films. In how many ways can we choose the 3 films if the order in which they are shown matters?

In this case, we have 5 choices for the first film, 4 choices for the second film (because one of the films has already been chosen), and 3 choices for the third film. This gives us

$$5 \cdot 4 \cdot 3 = 60$$

different ways of choosing the 3 films.

We can write the calculation in the example like this:

$$5 \cdot 4 \cdot 3 = \frac{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{2 \cdot 1} = \frac{5!}{2!} = \frac{5!}{(5-3)!} \; .$$

From this we deduce the following general formula:

Theorem 1.8

If we choose r elements out of n, we can do so in V_n^r ways if the order matters. The number V_n^r is given by this formula:

$$V_n^r = \frac{n!}{(n-r)!} \; .$$

In the cases, where $n = r$ (i.e. where we investigate in how many ways we can list all of the elements), we talk about *permutations*. The number of permutations of a set is the number of ways in which the set might be ordered.

Sometimes the term *permutations* is used about the number V_n^r, and the notation $P(n, r) = V_n^r$ is also used.

1.3 Combinations

If the order does not matter, we have fewer ways of choosing a number of elements of a larger set. If we want to choose 3 elements out of 5, we can still count our way to an answer if we approach the problem systematically. E.g. if we want to pick three letters from ABCDE, we get the possibilities in table 1.2. So, there are 10 different ways to choose 3 elements out of 5.

Table 1.2: The possibilities when choosing three letters from ABCDE.

ABC	ACD	BCD	CDE
ABD	ACE	BCE	
ABE	ADE	BDE	

poss

The reason why we have fewer possibilities is that when the order does not matter, then e.g. ABC and CBA will be the same choice. When the order *does* matter, we can choose 3 out of 5 in

$$\mathrm{V}_5^3 = \frac{5!}{(5-3)!} = 60$$

different ways. These 60 possibilities fall in groups of 6 containing the same 3 elements. This is because 3 elements can be *permuted* in $\mathrm{V}_3^3 = 6$ different ways. So, when the order *does not* matter, 3 out of 5 may be chosen in

$$\frac{\mathrm{V}_5^3}{\mathrm{V}_3^3} = \frac{5!}{(5-3)! \cdot 3!} = 10$$

different ways. We generalise this calculation to the following theorem:

Theorem 1.9

We can choose r elements out of n in C_n^r different ways if the order does not matter. The number C_n^r is given by

$$\mathrm{C}_n^r = \frac{n!}{r! \cdot (n-r)!} \,.$$

The number C_n^r is called the *binomial coefficient*.

A lot of different notation is used for the number of combinations. Besides C_n^r, e.g. $K(n,r)$ and $\binom{n}{r}$ are also used.

Example 1.10 In a deck of playing cards there are 52 cards. If we choose 5 of these, we can do so in

$$\mathrm{C}_{52}^5 = \frac{52!}{5! \cdot (52-5)!} = \frac{52!}{5! \cdot 47!} = 2\,598\,960$$

different ways.

2 Probability theory

As previously mentioned, probability deals with assigning numbers to random phenomena. The first books about probability dealt largely with games and gambling,[4] and the main concern was determining the probabilities of different *outcomes* of games.

An *outcome* is generally understood to mean the result of an "experiment", i.e. an event which can have several different results.

2.1 What is probability?

If we roll a dice, the probability of getting a 5 will be $\frac{1}{6}$, but what does that really mean? When we look at a regular six-sided dice, we assume that nothing makes one outcome more likely than another, and that we can find the probabilities by using this formula:

$$\text{probability} = \frac{\text{number of desired outcomes}}{\text{number of possible outcomes}}. \tag{2.1}$$

In this case, we have a so-called *a priori* probability, i.e. the probabilities are given prior to the experiment. We arrive at the value of the probability by deducing it from the experiment.

We can also talk about *frequentist* probabilities. This is the probability we get if we roll the dice a large number of times and calculate the relative frequency of 5s.

As noted, the formula above only applies to situations where all of the outcomes are equally likely. Therefore, it makes sense to try to describe the situations where we wish to find the probability, in such a way that the outcomes we describe are equally likely.

Table 2.1: Possible outcomes when rolling a dice.

s	p
⚀	$\frac{1}{6}$
⚁	$\frac{1}{6}$
⚂	$\frac{1}{6}$
⚃	$\frac{1}{6}$
⚄	$\frac{1}{6}$
⚅	$\frac{1}{6}$

2.2 Probability spaces

If we roll a dice, there are 6 possibilities for the result. These possibilities are listed in table 2.1. The 6 possibilities are equally likely, and they make up the *sample space* S which is the set of all possible outcomes,

$$S = \{⚀, ⚁, ⚂, ⚃, ⚄, ⚅\} \ .$$

The sum of the probabilities of all of the outcomes is 1. The sample space S and the related possibilities p are called a *probability space* (S, p). We have the following definition:

Definition 2.1

If $S = \{s_1, \dots, s_n\}$ is a set of outcomes, and $p_1, \dots, p_n$ are the related probabilities such that

1. the numbers $p_1, \dots, p_n$ are between 0 and 1, and
2. $p_1 + p_2 + \cdots p_n = 1$,

then (S, p) is called a finite[1]probability space.

[1]"Infinite" probability spaces also exist, in which the number of elements is infinite (e.g. "every whole number"), but they are beyond the scope of this description.

If we want to calculate the probabilities $p_1, \dots, p_n$, it is easier if the sample space is chosen in such a way that all of the outcomes are equally probable; in this case we talk about a *symmetric probability space*. The good thing about a symmetric probability space is that formula (2.1) applies. I.e. a symmetric probability space is defined in the following way:

Definition 2.2

If for a probability space (S, p) with n elements we have

$$P(s_1) = P(s_2) = \cdots = P(s_n) = \frac{1}{n} ,$$

the probability space (S, p) is called a *symmetric* probability space.

Any subset of the sample space, i.e. a set containing some of the outcomes in the sample space, is called an *event*. So, an event is a term used to describe which outcomes we look at in a given situation (corresponding to the "desired outcomes" in the formula 2.1).

Some possible events when we roll a dice are

$$\begin{aligned} E_1 &= \{⚅\} \\ E_2 &= \{⚀, ⚁\} \\ E_3 &= \{⚀, ⚂, ⚄\} \ . \end{aligned}$$

The event E_1 corresponds to getting a 6, E_2 corresponds to getting a 1 or a 2, and E_3 corresponds to the rolling an odd number of eyes. All of these events describe something that might happen when we roll a dice once.

Each event E has an associated probability $P(E)$. We can describe the distribution of probabilities by listing the outcomes and their probabilities in a table. The probabilities associated with the roll of a dice are listed in table 2.1.

We can calculate the probability of an event by adding the probabilities of the outcomes which make up the event. For the three events described above, we have

$$\begin{aligned} P(E_1) &= p_6 = \tfrac{1}{6} \\ P(E_2) &= p_1 + p_2 = \tfrac{1}{6} + \tfrac{1}{6} = \tfrac{1}{3} \\ P(E_3) &= p_1 + p_3 + p_5 = \tfrac{1}{6} + \tfrac{1}{6} + \tfrac{1}{6} = \tfrac{1}{2} \ . \end{aligned}$$

Here, p_1 is the probability to roll a 1, etc.

When we calculate probabilities, we sometimes need to look a two (or more) things happening at the same time. In these cases, we need to know whether the events are so-called *independent* event. We have the following:

Definition 2.3

If for two events A and B in a probability space (U, p) we have

$$P(\text{both } A \text{ and } B) = P(A) \cdot P(B) \ ,$$

the two events are called *independent*.

Here we give an example of two independent events, and two events that are not independent:

Example 2.4 If we roll a dice twice, the events

$$A : \text{ we get a 6 on the first roll}$$
$$B : \text{ we get a 2 on the second roll}$$

are independent events because the result of the first roll does not influence the result of the second roll.

We can find an example of events which are *not* independent by putting 5 black and 5 red balls in a jar, and then draw two of the balls. If we look at the events

$$C : \text{ we draw a black ball the first time}$$
$$D : \text{ we draw a red ball the second time} \ ,$$

it is clear that they are not independent because the probability of the second event depends on whether we drew a black or a red ball the first time.

We use the formula above when we calculate the probability that two events both happen. The probability that an event A *or* another event B happens can also be calculated – in those cases where the two events have no outcome in common (i.e. none of the outcomes in A may be part of B and vice versa). In those cases, we have

$$P(A \text{ or } B) = P(A) + P(B) \ .$$

Table 2.2: Outcomes in the roll of a dice, and the two random variables X and Y.

s	$P(s)$	X	Y
⚀	$\frac{1}{6}$	1	0
⚁	$\frac{1}{6}$	2	1
⚂	$\frac{1}{6}$	3	0
⚃	$\frac{1}{6}$	4	1
⚄	$\frac{1}{6}$	5	0
⚅	$\frac{1}{6}$	6	1

2.3 Random variables

The three events E_1, E_2 and E_3 above can also be described by so-called *random variables* which links a number to the events we look at (i.e. “getting a 6”, “getting a 1 or 2”, “getting an odd number”). The values of the random variables X and Y are shown in table 2.2.

The values of the random variables are chosen in such a way that they describe the event as wekk as possible – in this case we look at the eyes on the dice, or whether they are even or odd. The numbers can actually be chosen at random, but we will usually chose them so that they describe the event in the best possible way. With the random variables X and Y, the event E_1, E_2 and E_3 may also be written as

$$E_1 = \{X = 6\}\ , \quad E_2 = \{X \leq 2\}\ , \quad E_3 = \{Y = 0\}\ .$$

So, we can write the probabilities of the three events like this:

$$P(H_1) = P(X = 6)\ , \quad P(H_2) = P(X \leq 2) \quad \text{and} \quad P(H_3) = P(Y = 0)\ .$$

In these examples, the concept of random variables may seem like an unnecessary extra layer, because we can just count the probabilities. But as soon as the situation gets a little more complicated, the idea makes a lot more sense.

Example 2.5 If we roll two dice and look at the sum of the eyes, we can write a sample space containing the numbers 2, 3, …, 12. But these outcomes are not equally likely because we only get a 12 by rolling two 6s, but we can get 5 by rolling either a 1 and a 4, or a 2 and a 3. But because calculations are easier in a symmetric probability space, we instead analyse the situation in the following way:

Table 2.3: The possible values of X: the sum of the eyes when rolling two dice.

	⚀	⚁	⚂	⚃	⚄	⚅
⚀	2	3	4	5	6	7
⚁	3	4	5	6	7	8
⚂	4	5	6	7	8	9
⚃	5	6	7	8	9	10
⚄	6	7	8	9	10	11
⚅	7	8	9	10	11	12

If we distinguish between the two dice by letting one of them be white and the other black, the sample space S contains pairs (w, b) where w is the eyes of the white dice and b is the eyes of the black dice, In total, there are 36 of such pairs, so the sample space contains 36 elements:

S = {(⚀;⚀), (⚀;⚁), (⚀;⚂), (⚀;⚃), (⚀;⚄), (⚀;⚅),
(⚁;⚀), (⚁;⚁), (⚁;⚂), (⚁;⚃), (⚁;⚄), (⚁;⚅),
(⚂;⚀), (⚂;⚁), (⚂;⚂), (⚂;⚃), (⚂;⚄), (⚂;⚅),
(⚃;⚀), (⚃;⚁), (⚃;⚂), (⚃;⚃), (⚃;⚄), (⚃;⚅),
(⚄;⚀), (⚄;⚁), (⚄;⚂), (⚄;⚃), (⚄;⚄), (⚄;⚅),
(⚅;⚀), (⚅;⚁), (⚅;⚂), (⚅;⚃), (⚅;⚄), (⚅;⚅)}

Table 2.4: The probability distribution for the sum of the eyes when rolling two dice.

x	$P(X = x)$
2	$\frac{1}{36}$
3	$\frac{1}{18}$
4	$\frac{1}{12}$
5	$\frac{1}{9}$
6	$\frac{5}{36}$
7	$\frac{1}{6}$
8	$\frac{5}{36}$
9	$\frac{1}{9}$
10	$\frac{1}{12}$
11	$\frac{1}{18}$
12	$\frac{1}{36}$

All of these outcomes are equally likely, i.e. the probability of one of the outcomes is $\frac{1}{36}$. So, the probability of the pair (⚀;⚅) which corresponds to a white 1 and a black 6 is $\frac{1}{36}$.

We now define the random variable X whose value is the sum of the eyes of the two dice. The possible values of X are then the numbers from 2 to 12, and using table 2.3, we can write down the probability distribution of X.

E.g. if we want to know the probability of getting 9 by rolling two dice, we count the number of 9s in table 2.3 and multiply by $\frac{1}{36}$, which is the possibility of a single outcome. There are four 9s, so

$$P(X = 9) = 4 \cdot \frac{1}{36} = \frac{1}{9}\ .$$

The entire probability distribution of X is listed in table 2.4.

Example 2.6 If we toss a coin 3 times and count the number of "heads", we have 4 different possible outcomes; we can get "heads" 0, 1, 2 or 3 times. But these are not equally likely, so we should not chose them as outcomes.

Instead we define the outcomes to be the actual possible sequences when we toss a coin, i.e. combinations of "heads" and "tails":

ttt, tth, thh, etc.

If we look at the number of "heads" in 3 tosses of a coin, we have three outcomes which result in 2 "heads". The event which contains these three outcomes is

$$E = \{\text{hht}, \text{hth}, \text{thh}\} \ .$$

The sample space contains every possible outcome, i.e.

$$S = \{\text{hhh}, \text{hht}, \text{hth}, \text{thh}, \text{htt}, \text{tht}, \text{tth}, \text{ttt}\} \ .$$

Here we see that we have 8 equally likely outcomes.

We now let the random variable X be the number of "tails" in the three tosses, and make a table of the outcomes and the values of the random variable (see table 2.5). The event E mentioned above corresponds to $X = 2$.

We see in the table that $X = 2$ for 3 outcomes, i.e.

$$P(X = 2) = 3 \cdot \tfrac{1}{8} = \tfrac{3}{8} \ .$$

The entire probability distribution for the number of "heads" in 3 tosses of a coin is listed in table 2.6.

We can also shoe the probability distribution as a bar chart. A bar chart of the probability distribution in table 2.6 is shown in figure 2.7.

Table 2.5: The values of the random variable X for all possible outcomes of 3 tosses of a coin.

s	X
ppp	0
ppk	1
pkp	1
pkk	2
kpp	1
kpk	2
kkp	2
kkk	3

Table 2.6: X's probability distribution.

x	$P(X = x)$
0	$\frac{1}{8}$
1	$\frac{3}{8}$
2	$\frac{3}{8}$
3	$\frac{1}{8}$

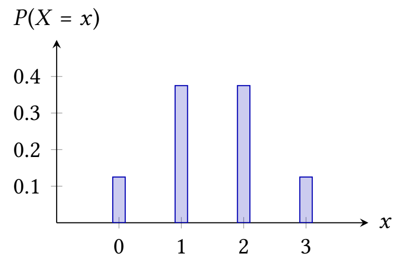


Figure 2.7: Bar chart of the probability distribution of X.

2.4 Mean and standard deviation

In many ways, probability theory is comparable to statistics, but instead of observations and relative frequencies, we talk about outcomes and probabilities. However, the methods are in many ways the same. E.g. probability distributions can be illustrated with bar charts as it was done in the example in the last section.

If we have a random variable and a probability distribution, we can therefore also calculate the mean, variance and standard deviation.

Definition 2.7

For a random variable X which can assume the values $x_1, \dots, x_n$, the mean μ, the variance σ^2, and the standard deviation σ is given by[2]

$$\mu = E(X) = \sum_{i=0}^{n} x_i \cdot P(X = x_i)$$

$$\sigma^2 = \mathrm{Var}(X) = \sum_{i=0}^{n} (x_i - \mu)^2 \cdot P(X = x_i)$$

$$\sigma = \sigma(X) = \sqrt{\mathrm{Var}(X)}\ .$$

[2]The E in $E(X)$ stands for "expectation value". The mean is also sometimes called the *expected value*.

We notice that the formulas are just like the ones used in statistics, but with the relative frequencies f_i replaced by the probabilities $P(X = x_i)$.

Example 2.8 If a random variable X counts the number of "heads" in 3 tosses of a coin, its probability distribution is given by table 2.6, and the mean is

$$\mu = 0 \cdot \frac{1}{8} + 1 \cdot \frac{3}{8} + 2 \cdot \frac{3}{8} + 3 \cdot \frac{1}{8} = 1.5\ ,$$

the variance is

$$\sigma^2 = (0 - 1.5)^2 \cdot \frac{1}{8} + (1 - 1.5)^2 \cdot \frac{3}{8} + (2 - 1.5)^2 \cdot \frac{3}{8} + (3 - 1.5)^2 \cdot \frac{1}{8} = \frac{3}{4}\ ,$$

and the standard deviation is

$$\sigma = \sqrt{\frac{3}{4}} = 0.866\ .$$

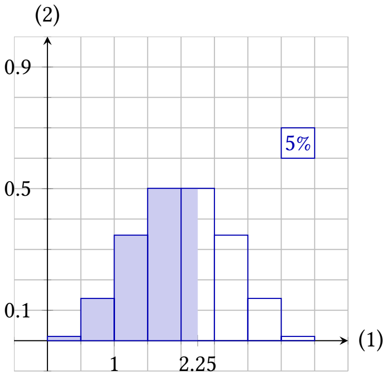
2.5 Discrete and continuous probabilities

If the sample space consists of a series of separate outcomes (like the ones we have previously looked at), the probability spaces is called *discrete*. When we roll a dice, we cannot get every number between 1 and 6, e.g. we cannot get 1.9 or 2.7. The sample space is therefore discrete. Another example would be a coin toss where we count the number of "heads". Here, we can get whole numbers, 1, 2, 3, …—but not e.g. 2.5. The methods we use when we look at discrete probability spaces are similar to use we use in ungrouped statistics.

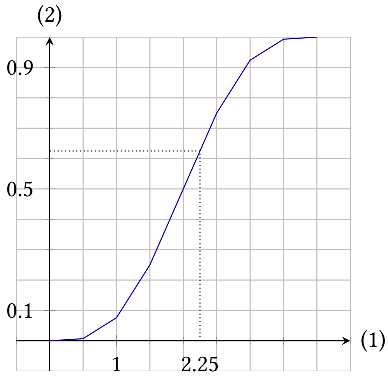
However, if the outcomes can be *any number* between some minimum and maximum values, we say that the probability space is *continuous*. Here we use methods that are similar to those used in grouped statistics. In this case, we calculate probabilities of intervals instead of probabilities of single values.

Example 2.9 During an autumn day, the minimum tempeture is 0°C, and the maximum temperature is 4°C. The temperature at a random time of day is then a continuous random variable, because the temperature can assume all values in the interval $[0; 4]$.

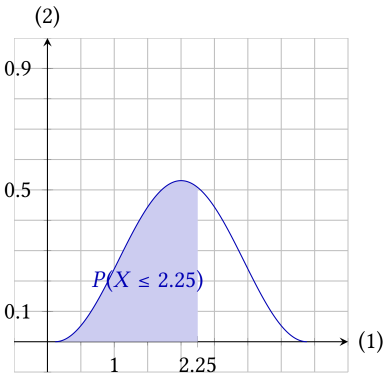
If we assume that we have measured the temperature continuously during the day, we can make a histogram which shows the distribution of temperatures, and a distribution curve which shows in what fraction of the day



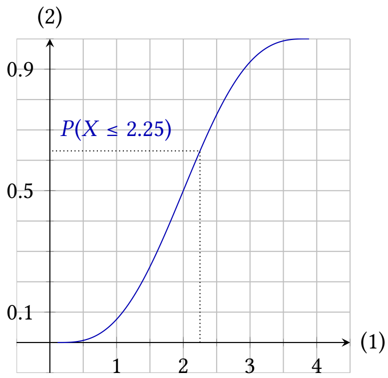
(a) Histogram.



(b) Distribution curve.



(c) Probability density function.



(d) Cumulative distribution function.

Figure 2.8: The relationship between histogram and distribution curve, and probability density function and cumulative distribution function.

the temperature was below a given value. The two diagrams might look like figure 2.8(a) and 2.8(b).

If we have measured the temperature continuously during the day, we might make the intervals for the histogram and the distribution curve smaller and smaller. We then imagine making the intervals "infinitely small" until we end up with the graphs of two continuous functions, see figure 2.8(c) and 2.8(d).

These two functions are called the *probability density function* and the *cumulative distribution function*—usually abbreviated PDF and CDF. Just as the area below a histogram equals the relative frequency of an interval, so does the area below the PDF equal the probability of an event, and just as the distribution curve is increasing from 0 to 1, so is the PDF.

As the example above shows, continuous random variables can be described by their probability density functions or their cumulative distribution functions. The probability that an outcome is less than a given value is then found be calculating the area below the graph of the PDF up to this value. But we can also find this probability directly from the graph of the CDF. If the CDF of a continuous random variable X is called F, then we have

$$P(X \leq a) = F(a) .$$

If we want to find the probability for an outcome in the interval $[a; b]$, we get

$$P(a \leq X \leq b) = P(X \leq b) - P(X \leq a) = F(b) - F(a) \ .$$

For continuous probability distributions we have:

Theorem 2.10

If X is a continuous random variable with cumulative distribution function F, then

1. $P(X \leq a) = F(a)$
2. $P(X \geq a) = 1 - F(a)$
3. $P(a \leq X \leq B) = F(b) - F(a)$.

A special case is the probability for a specific outcome. This is always 0 because

$$P(X = a) = P(a \leq X \leq a) = F(a) - F(a) = 0 \ .$$

We arrive at this (somewhat counter-intuitive) result because when a random variable is continuous its sample space contains an infinite amount of specific numbers. The probability of getting exactly one of those is then infinitely small, i.e. 0.

The calculus connection

Because we find the probabilities of events in a continuous probability space by determining areas under graphs, we can calculate these probabilities by integration. If we know integral calculus, we can define probability density functions like this:

Definition 2.11

A *probability density function* f is a function such that $f(x) \geq 0$ for all $x \in \mathbb{R}$, and

$$\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1 \ .$$

The probability of an event is then given by

Theorem 2.12

Let X be a continuous random variable with probability density function f, and let the interval $E = [e_1; e_2]$ be an event. Then the probability $P(E)$ of the event E is given by

$$P(E) = \int_{e_1}^{e_2} f(x)\,\mathrm{d}x \ .$$

Because the value of the cumulative distribution function is given by $F(a) = P(X \leq a)$, we define cumulative distribution functions in the following way:

Definition 2.13

Let X be a continuous random variable with probability density function f. Then the cumulative distribution function is given by

$$F(a) = \int_{-\infty}^{a} f(x)\,\mathrm{d}x\ .$$

For continuous random variables, we also calculate the mean and the variance as integrals. We have the following definition:

Definition 2.14

For a continuous random variable X with probability density function f, the mean μ, the variance σ^2, and the standard deviation σ are given by

$$\mu = E(X) = \int_{-\infty}^{\infty} x \cdot f(x)\,\mathrm{d}x$$

$$\sigma^2 = \mathrm{Var}(X) = \int_{-\infty}^{\infty} (x - \mu^2) \cdot f(x)\,\mathrm{d}x$$

$$\sigma = \sigma(X) = \sqrt{\mathrm{Var}(X)}\ .$$

The binomial distribution

3

The *binomial distribution* is a probability distribution which is used to calculate the probability of getting a certain amount of successes in a series of experiments. E.g. the probability of getting three 6s when rolling a dice five times.

When we want to calculate the probability of this outcome, we start with an experiment (called a trial) which we perform a number of times. Each time the trial is performed, we have a probability of success p, and a probability of failure $1 - p$.[1]

[1]Failure is the opposite success, so the probability of success and the probability of failure must add up to 1.

An example of a trial is the roll of a dice. The trial is then performed five times. If success is getting a 6, then the probability of success is $\frac{1}{6}$ (because the probability of rolling a 6 in one roll of a dice is $\frac{1}{6}$). The probability of failure is then $1 - \frac{1}{6} = \frac{5}{6}$ which corresponds to the probability of getting something else than a 6.

If we want to find the probability of rolling three 6s out of the five rolls, we first need to consider that the three 6s can be obtained in different ways. The first three rolls might be 6s, or it might be the last three rolls; there are a lot of different ways of getting three 6s out of five. Therefore, to determine the probability, we first need to determine in how many ways we can get three 6s in five rolls.

Here we can use the binomial coefficient (see theorem ??). Three 6s in five rolls can be obtained in $C_5^3 = 10$ different ways.

One of the possibilities is getting 6s in the first three rolls. The probability of getting one 6 is $\frac{1}{6}$. This has to happen the first three times. The fourth and the fifth roll cannot be 6s, the probability of this is $\frac{5}{6}$. So, the total probability of getting three 6s in the first three rolls and the something else in the last two is

$$\overbrace{\frac{1}{6} \cdot \frac{1}{6} \cdot \frac{1}{6}}^{\text{The first three}} \cdot \overbrace{\frac{5}{6} \cdot \frac{5}{6}}^{\text{The last two}} = \left(\frac{1}{6}\right)^3 \cdot \left(\frac{5}{6}\right)^2 .$$

All of the ways in which we can get three 6s must be equally likely. If we are interested in the probability of getting three 6s in five rolls (and not just getting 6s on the first three), we have to multiply this probability by the number of ways in which we can get three 6s, i.e. the probability becomes

$$10 \cdot \left(\frac{1}{6}\right)^3 \cdot \left(\frac{5}{6}\right)^2 = \frac{125}{3888} \approx 0.0322 . \tag{3.1}$$

If we gather all of the calculations into one, we can write (3.1) like this

$$C_5^3 \cdot \left(\frac{1}{6}\right)^3 \cdot \left(1 - \frac{1}{6}\right)^{5-3} . \tag{3.2}$$

Here, we are only using the original numbers: The number of 6s (3), the number of rolls (5), and the probability of getting a 6 in one roll ($\frac{1}{6}$).

3.1 The general formula

We want to write down a general formula for the binomial distribution, we define the random variable X which counts the number of successes in n trials. In each individual trial, the probability of success is p.

We then say that X is binomially distributed with *number of trials* n and *probability of success* p, and we write $X \sim b(n, p)$. The probability of r successes can the be calculated using the following formula which is a generalisation of the calculation (3.2).

Theorem 3.1

If the random variable X is binomially distributed with number of trials n and probability of success p, $X \sim b(n, p)$, the probability of r successes is given by

$$P(X = r) = C_n^r \cdot p^r \cdot (1 - p)^{n-r} .$$

Example 3.2 What is the probability of getting exactly four 1s in 15 rolls of a dice?

The random variable which counts the number of 1s in 15 rolls is binomially distributed with number of trials 15 and probability of success $\frac{1}{6}$, $X \sim b\left(15, \frac{1}{6}\right)$. Therefore, the probability of getting four 1s is

$$\begin{aligned} P(X = 4) &= C_{15}^4 \cdot \left(\frac{1}{6}\right)^4 \cdot \left(1 - \frac{1}{6}\right)^{15-4} \\ &= 1365 \cdot \left(\frac{1}{6}\right)^4 \cdot \left(\frac{5}{6}\right)^{11} = 0.1418 . \end{aligned}$$

So, there is a 14.18% chance of getting exactly four 1s in 15 rolls of a dice.

Example 3.3 On a little tropical island in the Pacific, there is a flood in the summer every four years on average. So, the probability of a flood occuring during a single summer is $\frac{1}{4}$. During a 5-year period, they can have everything from 0 to 5 floods. The probability distribution is then found by calculating $P(X = 0), P(X = 1), \ldots, P(X = 5)$.

E.g. we have

$$P(X = 3) = C_5^3 \cdot \left(\frac{1}{4}\right)^3 \cdot \left(\frac{3}{4}\right)^2 = 0.0879 .$$

So, this is the probability that a flood occurs 3 times during a 5-year period. The entire distribution is listed in table 3.1. A bar chart is shown in figure 3.2.

Table 3.1: The probability of r floods in a 5-year period.

r	$P(X = r)$
0	0.2373
1	0.3955
2	0.2637
3	0.0879
4	0.0146
5	0.0010

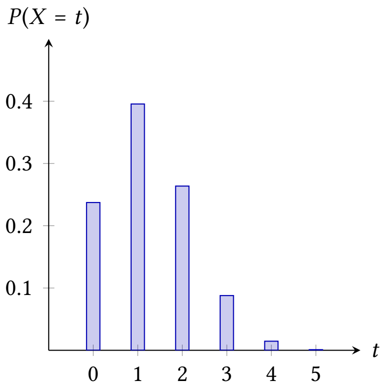


Figure 3.2: The probability distribution of X: Number of floods during a 5-year period.

As the table and the figure show, it is most likely to have a flood in one year; but we can also see that there is rather large probability of not getting any floods at all during the 5 years. However, it is very unlikely (0.0010) that they will have floods every year during a 5-year period.

If we want to know the probability of at most one flood during the 5 years, we have to calculate

$$P(X \leq 1) = P(X = 0) + P(X = 1) = 0.2373 + 0.3955 = 0.6328 \; .$$

So, it is very likely to get at most one flood during the 5 years. However, there is also a probability of

$$P(X > 1) = 1 - P(X \leq 1) = 1 - 0.6328 = 0.3672$$

of getting more than one flood during the 5 years.

3.2 Mean and standard deviation

The mean and standard deviation of a binomially distributed random variable are given by the following formulas which we will not prove:[1]

Theorem 3.4

If the random variable X is binomially distributed, $X \sim b(n, p)$, then the mean μ and the standard deviation σ are given by

$$\mu = np$$
$$\sigma = \sqrt{np(1-p)} \; .$$

Example 3.5 If we roll a dice 10 times and count the number of 5s, the random variable representing this number is binomially distributed with number of trials $n = 10$ and probability of success $p = \frac{1}{6}$.

The mean is then

$$\mu = n \cdot p = 10 \cdot \frac{1}{6} \approx 1.667 \; .$$

So, if we roll a dice ten times, we will get 1.667 5s on average.

The standard deviation is

$$\sigma = \sqrt{np(1-p)} = \sqrt{10 \cdot \frac{1}{6} \cdot \left(1 - \frac{1}{6}\right)} = 1.179 \; .$$

Example 3.6 In example 3.3, we looked at a random variable with number of trials 5 and probability of success $\frac{1}{4}$.

Here, the mean is

$$\mu = n \cdot p = 5 \cdot \frac{1}{4} = 1.25 \; ,$$

and the standard deviation is

$$\sigma = \sqrt{np(1-p)} = \sqrt{5 \cdot \frac{1}{4} \cdot \left(1 - \frac{1}{4}\right)} = 0.9682 \; .$$

3.3 Binomial test

James Bond is famous for wanting his martini "shaken not stirred". But can he really taste the difference? Assume that he drinks 16 martinis and has two tell how the martinis were made, and that he gets it right 13 times. How can we decide if this is an adequate number of right answers to know that he is not just guessing?

The experiment itself consists if 16 repetitions of the trial *decide whether the martini is "shaken" or "stirred"*. The binomial distribution is therefore at the core of out assessment, and the test we are about to perform is called a *binomial test.*

Before we do the test, we need to write down a so-called *null hypothesis* which will provide us with a probability of success. Here, we choose the null hypothesis

H_0: James Bond cannot taste the difference between the two types of martini.

We formulate the hypothesis in this way because this is what we can test. We know what the probabilities are if he is merely guessing (then the probability of success is $\frac{1}{2}$)—but we have no way of knowing what the probability of success is if he can actually tell the difference.

Next, we choose a so-called *significance level* which tells us when to reject the null hypothesis. A typical choice is 5%. In the test, we then try to investigate which results of the sample yield probabilities less than the significance level.

In our case, we investigate probabilities of the form

$$P(X \geq k) ,$$

where k is a whole number. E.g. we have

$$P(X \geq 11) = 0.105 = 10.5\%$$
$$P(X \geq 12) = 0.038 = 3.8\% .$$

Here, we see that there is less than 5% probability of getting 12 or more answers right. Therefore, so-called *critical region* is

$$\{12, 13, 14, 16\} .$$

The probability of this set is less then 5%, i.e. less than the chosen significance level.

Because James Bond's result lies in the critical region, there is less than 5% probability to see this result if he merely guesses. So, we choose to *reject* the null hypothesis. This means that we accept that he *can* taste how the martini was made.

The test we performed is a so-called *right-sided* test because we test whether his result is too large for him to just guess. In other situations, we might want to test whether a given value is too small; in these situations we use a *left-sided* test:

Right-sided test In a right-sided binomial test with significance level α, the critical region is

$$K = \{k, k+1, \dots, n\} \ ,$$

where k is the smallest number so that $P(X \leq k) \leq \alpha$.

Left-sided test In a left-sided binomial test with significance level α, the critical region is

$$K = \{0, 1, \dots, k\} \ ,$$

where k is the largest number so that $P(X \leq k) \leq \alpha$.

Example 3.7 A company producing canned tomatoes promise that 98% of the cans will arrive undamaged after delivery. A supermarket receives a pallet of tomatoes with 960 cans. 25 of the cans are damaged.

To find out whether we should trust the companies claims, we perform a left-sided binomial test. In this case, the null hypothesis is

H_0: 98% of the cans are undamaged.

The binomial distribution then has $n = 960$ and $p = 98\%$. At a significance level of 5%, we find

$$P(X \leq 923) = 0.0468 = 4.68\%$$
$$P(X \leq 924) = 0.0711 = 7.11\% \ .$$

So, the critical region is

$$K = \{1, 2, 3, \dots, 921, 922, 923\} \ .$$

935 of the cans are undamaged, and this number falls outside the critical region. Therefore, we accept the null hypothesis. At a 5% significance level, we can trust the company's claims.

In the test above, we tested whether James Bond can tell the difference between two kinds of martinis by investigating whether the number of right answers was large enough to reject the null hypothesis.

We can also do a *two-sided* test. Here we test whether the number is either too large or too small. If we reject the null hypothesis in this case, we conclude that he can tell the difference (but we cannot conclude that he knows which is which).

Because the test is two-sided, we divide the significance level by 2 (here we get 2.5%) and investigate when $P(X \leq a)$ and $P(X \geq b)$ are larger than 2.5%. We have

$$P(X \leq 3) = 0.0106 = 1.06\%$$
$$P(X \leq 4) = 0.0384 = 3.84\%$$

og

$$P(X \geq 12) = 0.0384 = 3.84\%$$

$$P(X \geq 13) = 0.0106 = 1.06\% \ .$$

In this case the critical region becomes

$$K = \{0, 1, 2, 3\} \cup \{13, 14, 15, 16\} \ .$$

Generally, we find the critical region in a two-sided binomial test like this:

Two-sided test In a two-sided binomial test with significance level α, the critical region is

$$K = \{0, 1, \ldots, a\} \cup \{b, \ldots, n\} \ ,$$

where a is the largest number so that $P(X \leq a) \leq \frac{\alpha}{2}$, and b is the smallest number so that $P(X \geq b) \leq \frac{\alpha}{2}$.

Example 3.8 In Waffleworth Heights, the Protest Party received 17.2% of the votes in the last communal election. In a opinion poll, where a representative sample of 1000 people were asked, 243 people say that they will vote for the party in the next election.

If we want to know whether the party's share of voters has changes, we perform a two-sided binomial test. The null hypothesis is then

$$H_0\text{: The percentage of votes for the party is 17.2\%.}$$

The binomial distribution then has $n = 1000$ and $p = 17.2\%$. If the significance level is $\alpha = 5\%$, then $\frac{\alpha}{2} = 2.5\%$. For the lower limit we find

$$\begin{aligned} P(X \leq 148) &= 0.0229 = 2.29\% \\ P(X \leq 149) &= 0.028 = 2.8\% \end{aligned}$$

and for the upper limit

$$\begin{aligned} P(X \geq 196) &= 0.0259 = 2.59\% \\ P(X \geq 197) &= 0.0214 = 2.14\% \ . \end{aligned}$$

From these calculations, we now see that the critical region is

$$K = \{0, 1, \ldots, 148\} \cup \{197, 198, \ldots, 1000\} \ .$$

Because the number 243 falls in the critical region, we reject the null hypothesis. So, the party's share of voters *has* changed since the election.

4 The normal distribution

Many statistical measurements result in frequency distributions which approximately follow the probability distribution we call the normal distribution. An example of this is the thickness of slices of bread cut on a machine.

No machine can cut perfectly. Table 4.1 lists measurements for a machine which is supposed to slice bread into slices with a thickness of 1 cm. Some of the slices are too thick and some are too thin; but most of the slices seem to have a thickness around 1 cm.

Table 4.1: Thickness of 100 slices of bread sliced on a machine.

Thickness (cm)	Number
0.55–0.65	2
0.65–0.75	4
0.75–0.85	6
0.85–0.95	9
0.95–1.05	14
1.05–1.15	16
1.15–1.25	15
1.25–1.35	10
1.35–1.45	10
1.45–1.55	8
1.55–1.65	6

Using the numbers in the table, we can find the mean $\overline{x}$ and the sample standard deviation s for the thickness of the slices. We get

$$\overline{x} = 1.151 \qquad \text{and} \qquad s = 0.249 \; .$$

Figure 4.2 shows a histogram of the distribution from table 4.1. The figure also shows the graph of the cumulative probabilities function of the *normal distribution with mean 1.51 and standard deviation 0.249*. The graph of the PDF is a bell-shaped curve which seems to fit quite nicely with the measured values of the thickness of the slices.

If a continuous random variable X is normally distributed with mean μ and standard deviation σ, we write $X \sim N(\mu, \sigma)$. If X is normally distributed, it has the following PDF:

Definition 4.1

A random variable X is called normally distributed with mean μ and standard deviation σ, $X \sim N(\mu, \sigma)$, if it has the probability density function

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \; .$$

Many phenomena result in normally distributed data. Some of them are:

- Measurement errors in experiments.
- The size of things that have been produced mechanically (like the thickness of the slices of bread above).
- Biological variables such as height and weight.[1]

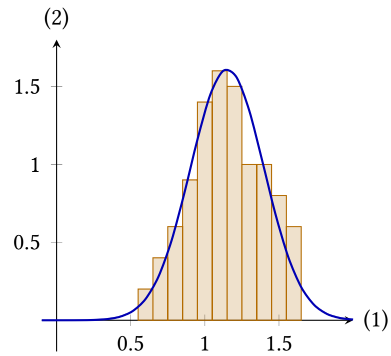
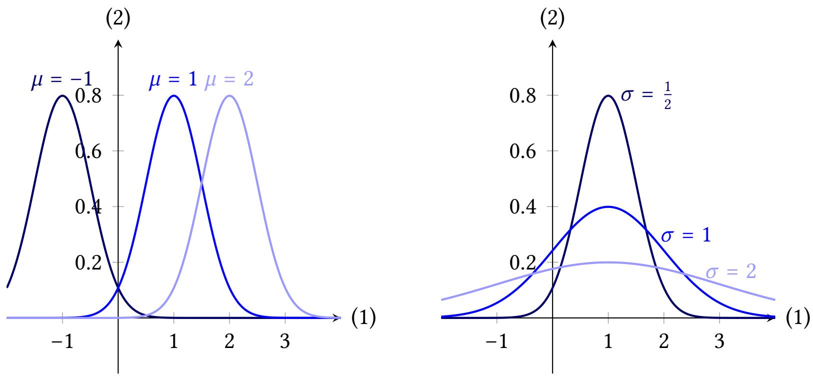


Figure 4.2: Histogram of the thickness of the slices.

The mean and the standard deviation change the shape of the curve. If you change the mean, the curve moves in a horizontal direction. If the standard

[1]However, many biological variables are only approximately normally distributed and are actually log-normally distributed[5].



(a) Different means.
(b) Different standard deviations.

Figure 4.3: If the PDFs have the same standard deviation but different means, the graphs are shifted horizontally. When they have the same mean but different standard deviations, the width of the curve changes.

deviation gets less, the curve becomes narrower; if the standard deviation gets larger, the curve will become wider (see figure 4.3).

As with every continuous probability distribution, we find the probability of a specific interval by determining the area under the graph of the PDF in the given interval.

Example 4.2 A machine at a factory fills 1 kg bags of sugar. The weight of the bags is normally distributed with mean $\mu = 1000$ g and standard deviation $\sigma = 25$ g. The PDF is then

$$f(x) = \frac{1}{\sqrt{2\pi} \cdot 25} \cdot \mathrm{e}^{-\frac{(x-1000)^2}{2 \cdot 25^2}} .$$

However, it is easier to calculate probabilities using the CDF, F.[2] E.g. the probability to take a sample and get a bag which weighs between 950 and 975 g is

$$\begin{aligned} P(950 \leq X \leq 975) &= F(975) - F(950) \\ &= 0.1359 = 13.59\% . \end{aligned}$$

[2] It is not possible to write an algebraic expression for the CDF, but it is a part of most CASs

So, the probability of getting a bag which weighs a bit less than 1 kg is not negligible.

Example 4.3 Here, we look again at the example with the slices of bread. It turned out that the random variable equal to the width of the slices was approximated quite well by a normal distribution with mean $\mu = 1.151$ and standard deviation $\sigma = 0.249$.

This allows us to answer questions like:

1. What is the probability of getting a slice of bread with a thickness between 0.9 cm and 1 cm?
2. What is the probability of getting a slice of bread with a thickness of more than 1.3 cm?

We can answer both by using the CDF to calculate the probabilities. The probability of getting slice of bread with a thickness between 0.9 cm and 1 cm is

$$P(0.9 \leq X \leq 1) = F(1) - F(0.9) = 0.1154 .$$

The probability of getting a slice of bread with a thickness of more than 1.3 cm is

$$P(X \geq 1.3) = 1 - P(X \leq 1.3) = 1 - F(1.3) = 0.2748 \ .$$

The areas which yield these probabilities are shown in figure 4.4.

The graph of the PDF of the normal distribution is symmetricl around the mean. The PDF is actually so well-behaved that we have the following theorem, which we will not prove here.[3]

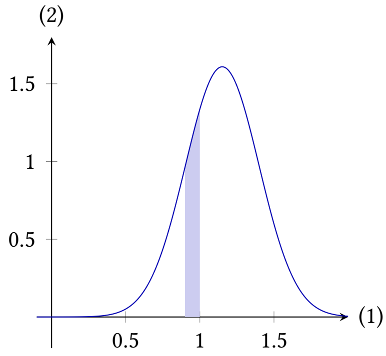
Theorem 4.4

If X is a normally distributed random variable with mean μ and standard deviation σ, then

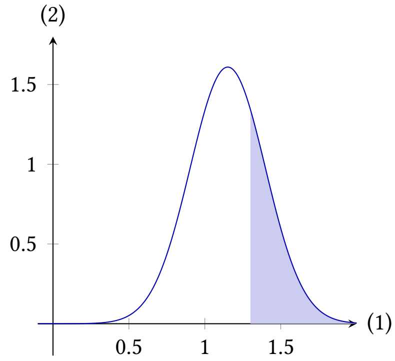
$$P(\mu - \sigma \leq X \leq \mu + \sigma) = 0.6827$$
$$P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = 0.9545$$
$$P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) = 0.9973 \ .$$

This theorem states that 68.27% of the outcomes will be in an interval of 1 standard deviation to each side of the mean, 95.45% will be in an interval 2 standard deviations to each side of the mean, etc. This is illustrated in figure 4.5.

Because most of the outcomes are in the interval $\mu \pm 2\sigma$, we call outcomes in this interval *normal outcomes*. Outcomes more than 3σ from the mean are called *exceptional outcomes*. These outcomes only make up $1 - 0.9973 = 0.0027 = 0.27\%$ of the entire distribution.



(a) $P(0.9 \leq X \leq 1) = 0.1154$.



(b) $P(X \geq 1.3) = 0.2748$.

Figure 4.4: The probabilities that the thickness of the slices fall in certain intervals are given by areas below the PDF.

4.1 Approximating the binomial distribution

If a random variable is binomically distributed $X \sim b(n, p)$, it turns out that we can approximate the distribution of X by a normal distribution with the same mean and standard deviation as the binomial distribution.

The approximation is very good when $\sigma^2 = np(1-p) \geq 10$, and it gets better, the larger this quantity gets.[1]

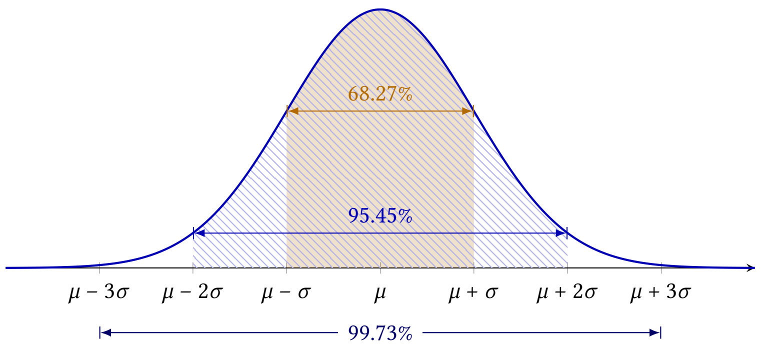


Figure 4.5: For a normally distributed random variabel, the probability that X is in a symmetric interval of 1 standard deviation to each side of the mean is a fixed number. So is the probability of an interval of 2 standard deviations to each side of the mean, etc.

Example 4.5 If a random variable $X \sim b(10, 0.4)$ is binomially distributed, its mean and standard deviation are

$$\mu = n \cdot p = 10 \cdot 0.4 = 4$$
$$\sigma = \sqrt{np(1-p)} = \sqrt{10 \cdot 0.4 \cdot 0.6} = \sqrt{2.4} = 1.55 \ .$$

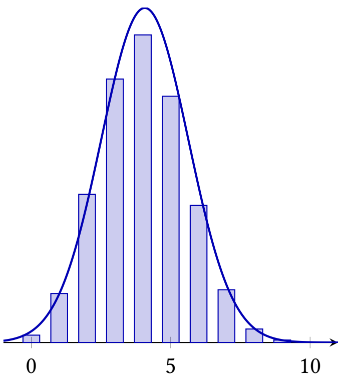


Figure 4.6: The distribution $b(10, 0.4)$ approximated by a normal distribution.

Here, σ^2 is 2.4 which is somewhat less than 10, i.e. we do not expect the normal distribution to be a very good approximation of this binomial distribution.

Figure 4.5 shows a bar chart of the probability distribution of X and the graph of the PDF af the normal distribution with mean 4 and standard deviation 1.55, and as we see the approximation is not that good.

However, if we look at a random variable Y which is binomially distributed $Y \sim b(50, 0.4)$, we see a better approximation. Here, the mean and the standard deviation are

$$\mu = 50 \cdot 0.4 = 20$$
$$\sigma = \sqrt{50 \cdot 0.4 \cdot 0.6} = \sqrt{12} = 3.46 \ ,$$

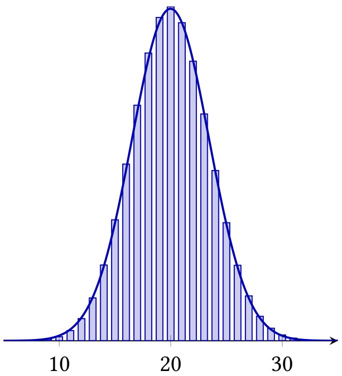


Figure 4.7: The distribution $b(50, 0.4)$ approximated by a normal distribution.

and σ^2 is 12, so a little more than 10. If we draw a bar chart of the distribution of Y and the PDF of the normal distribution with mean 20 and standard deviation 3.46, we get figure 4.5.

Here, we see a clear similarity between the graph and the bar chart, i.e. this normal distribution is a good approximation of the binomial distribution.

4.2 Samples

Because we can use the normal distribution to approximate the normal distribution, we can also use it to determine a *confidence interval* for the probability of success p in a binomial distribution. To determine this interval we first need an estimate of the parameter p.

Example 4.6 A company wants to investigate how many percent of consumers know about a new product. They conduct a survey and ask 1142 representatively chosen people whether they have heard of this new product. Out of the 1142, 715 have heard of the new product.

Based on these numbers, they estimate that

$$\hat{p} = \frac{715}{1142} = 0.626 = 62.6\%$$

of consumers have heard of the new product.

How sure can we be of this number? If we assume that $\hat{p}$ is a good estimate of the true parameter p, then the number of people who have heard about the product is binomially distributed with number of trials $n = 1142$ and probability of success $\hat{p} = 0.626$. The mean of this binomial distribution is

$$\mu = n\hat{p} \ ,$$

and the standard deviation is

$$\sigma = \sqrt{n\hat{p}(1-\hat{p})} \ .$$

Now, if we approximate this binomial distribution by a normal distribution, we find that 95.45% of the distribution is in the interval

$$\mu \pm 2\sigma = n\hat{p} \pm 2 \cdot \sqrt{n\hat{p}(1-\hat{p})} \ .$$

Because we are more interested in a percentage than in the total number, we divide this number by n and get the interval

$$\hat{p} \pm 2 \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \ .$$

We can now say that with a probability of 95.45% the true value of the probability of success p will be in this interval.

Theorem 4.7

In a sample of n elements and n_s successes, the 95% confidence interval is given by

$$\left[\hat{p} - 2 \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} ; \hat{p} + 2 \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right] ,$$

where $\hat{p} = \frac{n_s}{n}$.

The interval given by the above formula is actually the 95.45% confidence interval for p. If we want exactly 95%, the 2 in the formula must be exchanged by 1.96.[3]

Example 4.8 In the previous example, we estimated the probability of success as $\hat{p} = 0.626$, i.e.

$$2 \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 2 \cdot \sqrt{\frac{0.626 \cdot (1-0.626)}{1142}} = 0.029 \ .$$

So, there is a 95% probability that the parameter p is in the interval

$$0.629 \pm 0.029$$

which means that with a probability of 95% between 59.7% and 65.5% of consumers have heard of the new product.

4.3 The standard normal distribution

If we do calculations involving the normal distribution, we would normally use a CAS. Before these were invented, tables of function values were used.

Because it is not possible to make tables for each possible value of the mean and standard deviation, tables were made of the so-called *standard normal distribution*. It turns out that this is sufficient because the PDF and the CDF of any normal distribution can be expressed via the standard normal distribution.

The standard normal distribution is defined in the following way:

Definition 4.9

The normal distribution with mean $\mu = 0$ and stanard deviation $\sigma = 1$ is called the *standard normal distribution*. The probability density function of this distribution is

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}x^2} .$$

The cumulative distribution function is

$$\Phi(x) = \int_{-\infty}^{x} \phi(t)\, dt .$$

The next few theorems show the connection between the PDF and the CDF of an arbitrary normal distribution and the corresponding functions for the standard normal distribution.

Theorem 4.10

For the probability density function $f(x)$ of the normal distribution with mean μ and standard deviation σ, we have

$$f(x) = \frac{1}{\sigma} \cdot \phi\left(\frac{x-\mu}{\sigma}\right) ,$$

where $\phi(x)$ is the probability density function of the standard normal distribution.

Proof
We prove the theorem by calculating $\frac{1}{\sigma} \cdot \phi\left(\frac{x-\mu}{\sigma}\right)$:

$$\begin{aligned} \frac{1}{\sigma} \cdot \phi\left(\frac{x-\mu}{\sigma}\right) &= \frac{1}{\sigma} \cdot \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \\ &= \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} = f(x) . \end{aligned}$$ ■

However, the tabulated function was not the PDF but the CDF, because this can be used for direct calculations of probabilities. We have the following theorem:

Theorem 4.11

For the cumulative distribution function $F(x)$ and the corresponding function $\Phi(x)$ for the standard normal distribution, we have

$$F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right) .$$

Proof
The cumulative distribution function of the normal distribution with mean μ and standard deviation σ is, according to theorem 4.10, equal to

$$F(x) = \int_{-\infty}^{x} f(t)\, dt = \int_{-\infty}^{x} \frac{1}{\sigma} \cdot \phi\left(\frac{t-\mu}{\sigma}\right) dt . \tag{4.1}$$

We now perform the substitution $u = \frac{t-\mu}{\sigma}$. Then $\mathrm{d}u = \frac{1}{\sigma} \cdot \mathrm{d}x$, and the expression in (4.1) becomes

$$\int_{-\infty}^{\frac{x-\mu}{\sigma}} \phi(u)\,\mathrm{d}u = \Phi\left(\frac{x-\mu}{\sigma}\right) .$$

So,

$$F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right) . \qquad \blacksquare$$

We do not need these theorems to find probabilities for a given normal distribution. But as it turns out, the standard normal distribution is quite useful when we want to investigate if a given data set is normally distributed.

4.4 Normally distributed data

In this section, we will show how to construct a so-called *quantile plot* or *QQ-plot* which is a graph that is used to determine whether a set of data is normally distributed.

Example 4.12 We want to know whether the following numbers are normally distributed (the numbers show the weights of 30 sacks of carrots which should each weigh 25 kg):

24.8	25.4	25.0	25.4	24.0	24.4
24.5	24.5	24.8	24.9	24.9	25.0
24.7	24.3	24.7	24.8	25.0	25.1
25.1	24.5	24.3	25.0	25.3	25.1
24.5	24.7	25.3	24.6	24.5	24.8

Table 4.8: The weight of 30 sacks of carrots, sorted and numbered.

Vægt. x	i	$z = \Phi^{-1}\left(\frac{i-0.5}{n}\right)$
24.0	1	−2.13
24.3	2	−1.64
24.3	3	−1.38
24.4	4	−1.19
⋮	⋮	⋮
25.3	28	1.38
25.4	29	1.64
25.4	30	2.13

If we calculate the mean and the sample standard deviation of this data set, we get

$$\overline{x} = 24.8 \quad \text{and} \quad s = 0.35 .$$

What we want to know is whether this data set corresponds to a normal distribution with this mean and standard deviation. First we sort the 30 numbers and number them (see table 4.8).

The idea behind the quantile plot is to compare the cumulative relative frequencies to the cumulative relative frequencies of the standard normal distribution. If the data is normally distributed, the cumulative relative frequencies will be the function values of a CDF $F(x)$ of a normal distribution with mean $\overline{x}$ and standard deviation s where

$$F(x) = \Phi\left(\frac{x-\overline{x}}{s}\right) ,$$

which we can rewrite as[3]

$$\Phi^{-1}(F) = \frac{x-\overline{x}}{s} .$$

This means that $\Phi^{-1}(F)$ is a linear function of x.

[3] Here, Φ^{-1} is the inverse function of the CDF of the standard normal distribution. This function can be found in most CASs.

Here, F is the cumulative relative frequency. In this example, the cumulative relative frequencies $F = \frac{i}{30}$ where i is measurement number i. So, the first measurement has the cumulative relative frequency $\frac{1}{30}$, and the last measurement has the cumulative relative frequency $\frac{30}{30} = 1$. This is a problem because the function $\Phi^{-1}(F)$ is not defined for $F = 1$.[4] So, in order to include the last point, we use the number $\frac{i-0,5}{n}$ (where n is the number of measurements, i.e. here $n = 30$) instead of the actual cumulative relative frequencies[2] and calculate the number z,

$$z = \Phi^{-1}\left(\frac{i-0,5}{n}\right) .$$

[4] This is because the CDF of the normal distribution will never have the function value 1, but only approach 1 as $x \longrightarrow \infty$.

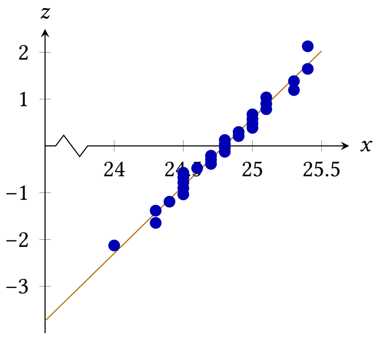


Figure 4.9: Quartile plot of the weight of carrots.

Next, we graph z as a function of x, see figure 4.9.

If the data set is approximately normally distributed, the points have to be approximately on the straight line $z = \frac{x-\overline{x}}{s}$ which in this case is

$$z = \frac{x - 24.9}{0.35} .$$

As we can see, the points are very close to this straight line, which means that this data set is normally distributed.

This is a brief summary of the method:

1. Calculate the mean $\overline{x}$ and the sample standard deviation s.
2. Make a table of the data set where the numbers are sorted and numbered.
3. Add a column containing $z = \Phi^{-1}\left(\frac{i-0,5}{n}\right)$.
4. Graph the points (x, z) in a coordinate system. x is the measurement.
5. Draw the line $z = \frac{x-\overline{x}}{s}$.
6. If the data set is normally distributed, the points are approximately on this line.

Luckily many CASs can make quantile plots so that we do not have to do them manually like this.

Bibliography

[1] Richard F. Bass et al. *Upper level undergraduate probability with actuarial and financial applications.* University of Connecticut, Department of Mathematics, 2018.

[2] Nicholas J. Cox. *Calculating percentile ranks or plotting positions.* URL: https://www.stata.com/ (visited on 03/23/2020).

[3] Sue Gordon. *The Normal Distribution.* Mathematics Learning Centre, 2006.

[4] Victor J. Katz. *A History of Mathematics. An Introduction.* 2nd ed. Addison-Wesley Educational Publishers, Inc., 1998.

[5] Eckhard Limpert, Werner A. Stahel, and Markus Abbt. “Log-normal Distributions across the Sciences: Keys and Clues”. In: *BioScience* 51.5 (2001), pp. 341–352.