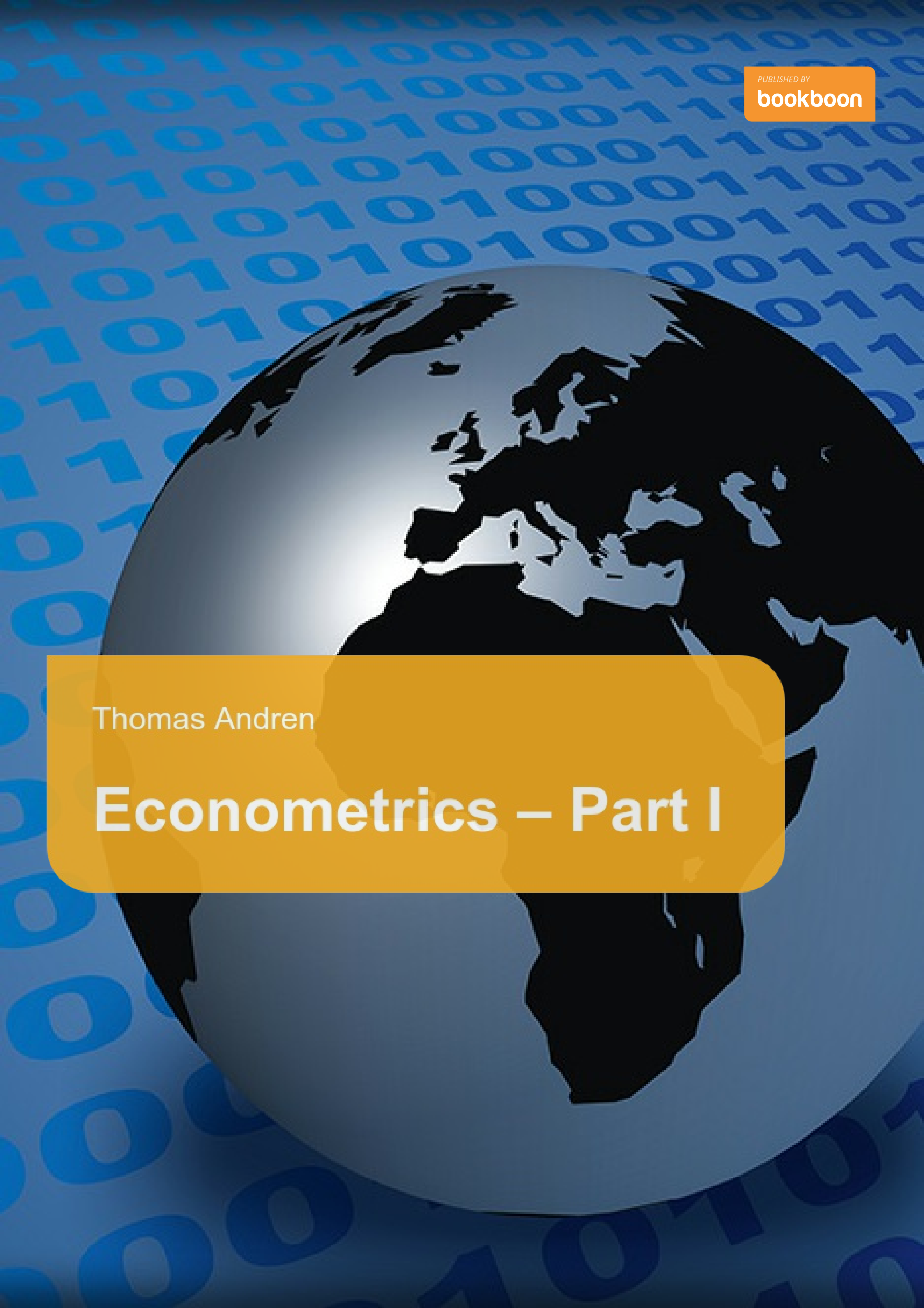
PUBLISHED BY bookboon



Thomas Andren

Econometrics – Part I

Thomas Andren

Econometrics

Part I

Econometrics – Part I
1st edition

ISBN 978-87-7681-315-4

Contents

1 Basics of probability and statistics

The purpose of this and the following chapter is to briefly go through the most basic concepts in probability theory and statistics that are important for you to understand. If these concepts are new to you, you should make sure that you have an intuitive feeling of their meaning before you move on to the following chapters in this book.

1.1 Random variables and probability distributions

The first important concept of statistics is that of a **random experiment**. It is referred to as any process of measurement that has more than one outcome and for which there is uncertainty about the result of the experiment. That is, the outcome of the experiment can not be predicted with certainty. Picking a card from a deck of cards, tossing a coin, or throwing a die, are all examples of basic experiments.

The set of all possible outcomes of on experiment is called the **sample space** of the experiment. In case of tossing a coin, the sample space would consist of a head and a tail. If the experiment was to pick a card from a deck of cards, the sample space would be all the different cards in a particular deck. Each outcome of the sample space is called a **sample point**.

An **event** is a collection of outcomes that resulted from a repeated experiment under the same condition. Two events would be **mutually exclusive** if the occurrence of one event precludes the occurrence of the other event at the same time. Alternatively, two events that have no outcomes in common are mutually exclusive. For example, if you were to roll a pair of dice, the event of rolling a 6 and of rolling a double have the outcome (3,3) in common. These two events are therefore not mutually exclusive.

Events are said to be **collectively exhaustive** if they exhaust all possible outcomes of an experiment. For example, when rolling a die, the outcomes 1, 2, 3, 4, 5, and 6 are collectively exhaustive, because they encompass the entire range of possible outcomes. Hence, the set of all possible die rolls is both mutually exclusive and collectively exhaustive. The outcomes 1 and 3 are mutually exclusive but not collectively exhaustive, and the outcomes even and not-6 are collectively exhaustive but not mutually exclusive.

Even though the outcomes of any experiment can be described verbally, such as described above, it would be much easier if the results of all experiments could be described numerically. For that purpose we introduce the concept of a random variable. A **random variable** is a function, which assigns unique numerical values to all possible outcomes of a random experiment.

By convention, random variables are denoted by capital letters, such as *X*, *Y*, *Z*, etc., and the values taken by the random variables are denoted by the corresponding small letters *x*, *y*, *z*, etc. A random variable from an experiment can either be **discrete** or **continuous**. A random variable is discrete if it can assume only a finite number of numerical values. That is, the result in a test with 10 questions can be 0, 1, 2, …, 10. In this case the discrete random variable would represent the test result. Other examples could be the number of household members, or the number of sold copy machines a given day.

Whenever we talk about random variables expressed in units we have a discrete random variable. However, when the number of unites can be very large, the distinction between a discrete and a continuous variable become vague, and it can be unclear whether it is discrete or continuous.

A random variable is said to be continuous when it can assume any value in an interval. In theory that would imply an infinite number of values. But in practice that does not work out. Time is a variable that can be measured in very small units and go on for a very long time and is therefore a continuous variable. Variables related to time, such as age is therefore also considered to be a continuous variable. Economic variables such as GDP, money supply or government spending are measured in units of the local currency, so in some sense one could see them as discrete random variables. However, the values are usually very large so counting each Euro or dollar would serve no purpose. It is therefore more convenient to assume that these measures can take any real number, which therefore makes them continuous.

Since the value of a random variable is unknown until the experiment has taken place, a probability of its occurrence can be attached to it. In order to measure a probability for a given events, the following formula may be used:

$$P(A) = \frac{\text{The number of ways event } A \text{ can occur}}{\text{The total number of possible outcomes}} \tag{1.1}$$

This formula is valid if an experiment can result in *n* mutually exclusive and equally likely outcomes, and if *m* of these outcomes are favorable to event *A*. Hence, the corresponding probability is calculated as the ratio of the two measures: n/m as stated in the formula. This formula follows the **classical definition** of a probability.

Example 1.1

You would like to know the probability of receiving a 6 when you toss a die. The sample space for a die is {1, 2, 3, 4, 5, 6}, so the total number of possible outcome are 6. You are interested in one of them, namely 6. Hence the corresponding probability equals 1/6.

Example 1.2

You would like to know the probability of receiving 7 when rolling two dice. First we have to find the total number of unique outcomes using two dice. By forming all possible combinations of pairs we have (1,1), (1,2),…, (5,6),(6,6), which sum to 36 unique outcomes. How many of them sum to 7? We have (1,6), (2,5), (3,4), (4,3), (5,2), (6,1): which sums to 6 combinations. Hence, the corresponding probability would therefore be 6/36 = 1/6.

The classical definition requires that the sample space is finite and that the each outcome in the sample space is equally likely to appear. Those requirements are sometimes difficult to stand up to. We therefore need a more flexible definition that handles those cases. Such a definition is the so called **relative frequency definition of probability** or the empirical definition. Formally, if in n trials, m of them are favorable to the event A, then $P(A)$ is the ratio m/n as n goes to infinity or in practice we say that it has to be sufficiently large.

Example 1.3

Let us say that we would like to know the probability to receive 7 when rolling two dice, but we do not know if our two dice are fair. That is, we do not know if the outcome for each die is equally likely. We could then perform an experiment where we toss two dice repeatedly, and calculate the relative frequency. In Table 1.1 we report the results for the sum from 2 to 7 for different number of trials.

	Number of trials						
Sum	10	100	1000	10000	100000	1000000	∞
2	0	0.02	0.021	0.0274	0.0283	0.0278	0.02778
3	0.1	0.02	0.046	0.0475	0.0565	0.0555	0.05556
4	0.1	0.07	0.09	0.0779	0.0831	0.0838	0.08333
5	0.2	0.12	0.114	0.1154	0.1105	0.1114	0.11111
6	0.1	0.17	0.15	0.1389	0.1359	0.1381	0.13889
7	0.2	0.17	0.15	0.1411	0.1658	0.1669	0.16667

Table 1.1 Relative frequencies for different number of trials

From Table 1.1 we receive a picture of how many trials we need to be able to say that that the number of trials is sufficiently large. For this particular experiment 1 million trials would be sufficient to receive a correct measure to the third decimal point. It seem like our two dices are fair since the corresponding probabilities converges to those represented by a fair die.

1.1.1 Properties of probabilities

When working with probabilities it is important to understand some of its most basic properties. Below we will shortly discuss the most basic properties.

1. $0 \leq \mathrm{P}(A) \leq 1$ A probability can never be larger than 1 or smaller than 0 by definition.
2. If the events *A*, *B*, … are mutually exclusive we have that $P(A+B+...)=P(A)+P(B)+...$

Example 1.4

Assume picking a card randomly from a deck of cards. The event *A* represents receiving a club, and event *B* represents receiving a spade. These two events are mutually exclusive. Therefore the probability of the event $C = A + B$ that represent receiving a black card can be formed by $P(A+B)=P(A)+P(B)$

3. If the events *A*, *B*, … are mutually exclusive and collectively exhaustive set of events then we have that $P(A+B+...)=P(A)+P(B)+...=1$

Example 1.5

Assume picking a card from a deck of cards. The event *A* represents picking a black card and event *B* represents picking a red card. These two events are mutually exclusive and collectively exhaustive. Therefore $P(A+B)=P(A)+P(B)=1$.

4. If event *A* and *B* are statistically independent then $P(AB)=P(A)P(B)$ where $P(AB)$ is called a joint probability.
5. If event *A* and *B* are not mutually exclusive then $P(A+B)=P(A)+P(B)-P(AB)$

Example 1.6

Assume that we carry out a survey asking people if they have read two newspapers (*A* and *B*) a given day. Some have read paper *A* only, some have read paper *B* only and some have read both *A* and *B*. In order to calculate the probability that a randomly chosen individual has read newspaper *A* and/or *B* we must understand that the two events are not mutually exclusive since some individuals have read both papers. Therefore $P(A+B)=P(A)+P(B)-P(AB)$. Only if it had been an impossibility to have read both papers the two events would have been mutually exclusive.

Suppose that we would like to know the probability that event *A* occurs given that event *B* has already occurred. We must then ask if event *B* has any influence on event *A* or if event *A* and *B* are independent. If there is a dependency we might be interested in how this affects the probability of event *A* to occur. The **conditional probability** of event *A* given event *B* is computed using the formula:

$$P(A \mid B)=\frac{P(AB)}{P(B)} \tag{1.2}$$

Example 1.7

We are interested in smoking habits in a population and carry out the following survey. We ask 100 people whether they are a smoker or not. The results are shown in Table 1.2.

	Yes	No	Total
Male	19	41	60
Female	12	28	40
Total	31	69	100

Table 1.2 A survey on smoking

Using the information in the survey we may now answer the following questions:

i) What is the probability of a randomly selected individual being a male who smokes?

This is just the joint probability. Using the classical definition start by asking how large the sample space is: 100. Thereafter we have to find the number of smoking males: 19. The corresponding probability is therefore: 19/100=0.19.

ii) What is the probability that a randomly selected smoker is a male?

In this case we focus on smokers. We can therefore say that we condition on smokers when we ask for the probability of being a male in that group. In order to answer the question we use the conditional probability formula (1.2). First we need the joint probability of being a smoker and a male. That turned out to be 0.19 according to the calculations above. Secondly, we have to find the probability of being a smoker. Since 31 individuals were smokers out of the 100 individuals that we asked, the probability of being a smoker must therefore be 31/100=0.31. We can now calculate the conditional probability. We have 0.19/0.31=0.6129. Hence there is 61% chance that a randomly selected smoker is a man.

1.1.2 The probability function – the discrete case

In this section we will derive what is called the **probability mass function** or just **probability function** for a stochastic discrete random variable. Using the probability function we may form the corresponding **probability distribution**. By probability distribution for a random variable we mean the possible values taken by that variable and the probabilities of occurrence of those values. Let us take an example to illustrate the meaning of those concepts.

Example 1.8

Consider a simple experiment where we toss a coin three times. Each trial of the experiment results in an outcome. The following 8 outcomes represent the sample space for this experiment: (HHH), (HHT), (HTH), (HTT), (THH), (THT), (TTH), (TTT). Observe that each sample point is equally likely to occure, so that the probability that one of them occure is 1/8.

The random variable we are interested in is the number of heads received on one trial. We denote this random variable *X*. *X* can therefore take the following values 0, 1, 2, 3, and the probabilities of occurrence differ among the alternatives. The table of probabilities for each value of the random variable is referred to as the probability distribution. Using the classical definition of probabilities we receive the following probability distribution.

X	0	1	2	3
P(*X*)	1/8	3/8	3/8	1/8

Table 1.3 Probability distribution for *X*

From Table 1.3 you can read that the probability that $X = 0$, which is denoted $P(X = 0)$, equals 1/8.

1.1.3 The cumulative probability function – the discrete case

Related to the probability mass function of a discrete random variable *X*, is its Cumulative Distribution Function, F(*X*), usually denoted CDF. It is defined in the following way:

$$F(X) = P(X \le c) \tag{1.3}$$

Example 1.9

Consider the random variable and the probability distribution given in Example 1.8. Using that information we may form the cumulative distribution for *X*:

X	0	1	2	3
P(*X*)	1/8	4/8	7/8	1

Table 1.4 Cumulative distribution for *X*

The important thing to remember is that the outcomes in Table 1.3 are mutually exclusive. Hence, when calculating the probabilities according to the cumulative probability function, we simply sum over the probability mass functions. As an example:

$$P(X \le 2) = P(X = 0) + P(X = 1) + P(X = 2)$$

1.1.4 The probability function – the continuous case

When the random variable is continuous it is no longer interesting to measure the probability of a specific value since its corresponding probability is zero. Hence, when working with continuous random variables, we are concerned with probabilities that the random variable takes values within a certain interval. Formally we may express the probability in the following way:

$$P(a \le X \le b) = \int_a^b f(x)dx \tag{1.4}$$

In order to find the probability, we need to integrate over the probability function, f(*X*), which is called the **probability density function**, pdf, for a continuous random variable. There exist a number of standard probability functions, but the single most common one is related to the standard normal random variable.

Example 1.10

Assume that *X* is a continuous random variable with the following probability function:

$$f(X) = \begin{cases} 3e^{-3X} & X > 0 \\ 0 & else \end{cases}$$

Find the probability $P(0 \leq X \leq 0.5)$. Using integral calculus we find that

$$P(0 \leq X \leq 0.5) = \int_0^{0.5} 3e^{-3x} dx = \left[-e^{-3x}\right]_0^{0.5} = \left[-e^{-3\times 0.5}\right] - \left[-e^{-3\times 0}\right] = -e^{-1.5} + 1 = 0.777$$

1.1.5 The cumulative probability function – the continuous case

Associated with the probability density function of a continuous random variable *X* is its **cumulative distribution function** (CDF). It is denoted in the same way as for the discrete random variable. However, for the continuous random variable we have to integrate from minus infinity up to the chosen value, that is:

$$F(c) = P(X \leq c) = \int_{-\infty}^{c} f(X)dX \tag{1.5}$$

The following properties should be noted:

1. $F(-\infty) = 0$ and $F(\infty) = 1$, which represents the left and right limit of the CDF.
2. $P(X \geq a) = 1 - F(a)$
3. $P(a \leq X \leq b) = F(b) - F(a)$

In order to evaluate this kind of problems we typically use standard tables, which are located in the appendix.

1.2 The multivariate probability distribution function

Until now we have been looking at univariate probability distribution functions, that is, probability functions related to one single variable. Often we may be interested in probability statements for several random variables jointly. In those cases it is necessary to introduce the concept of a **multivariate probability function**, or a **joint distribution function**.

In the discrete case we talk about the joint probability mass function expressed as

$$f(X,Y) = P(X = x, Y = y)$$

Example 1.11

Two people A and B both flip coin twice. We form the random variables *X* = "number of heads obtained by A", and *Y* = "number of heads obtained by B". We will start by deriving the corresponding probability mass function using the classical definition of a probability. The sample space for person A and B is the same and equals {(H,H), (H,T), (T,H), (T,T)} for each of them. This means that the sample space consists of 16 (4×4) sample points. Counting the different combinations, we end up with the results presented in Table 1.5.

		X			
		0	1	2	Total
Y	0	1/16	2/16	1/16	4/16
	1	2/16	4/16	2/16	8/16
	2	1/16	2/16	1/16	4/16
	Total	4/16	8/16	4/16	1.00

Table 1.5 Joint probability mass function, *f*(*X*, *Y*)

As an example, we can read that $P(X=0, Y=1)=2/16=1/8$. Using this table we can determine the following probabilities:

$$P(X<Y)=P(X=0,Y=1)+P(X=0,Y=2)+P(X=1,Y=2)=\frac{2}{16}+\frac{1}{16}+\frac{2}{16}=\frac{5}{16}$$

$$P(X>Y)=P(X=1,Y=0)+P(X=2,Y=0)+P(X=2,Y=1)=\frac{2}{16}+\frac{1}{16}+\frac{2}{16}=\frac{5}{16}$$

$$P(X=Y)=P(X=0,Y=0)+P(X=1,Y=1)+P(X=2,Y=2)=\frac{1}{16}+\frac{4}{16}+\frac{1}{16}=\frac{6}{16}$$

Using the joint probability mass function we may derive the corresponding univariate probability mass function. When that is done using a joint distribution function we call it the marginal probability function. It is possible to derive a marginal probability function for each variable in the joint probability function. The marginal probability functions for X and Y is

$$f(X) = \sum_{y} f(X,Y) \text{ for all } X \tag{1.6}$$

$$f(Y) = \sum_{x} f(X,Y) \text{ for all } Y \tag{1.7}$$

Example 1.12

Find the marginal probability functions for the random variables X.

$$P(X=0) = f(X=0,Y=0) + f(X=0,Y=1) + f(X=0,Y=2) = \frac{1}{16} + \frac{2}{16} + \frac{1}{16} = \frac{4}{16} = \frac{1}{4}$$

$$P(X=1) = f(X=1,Y=0) + f(X=1,Y=1) + f(X=1,Y=2) = \frac{2}{16} + \frac{4}{16} + \frac{2}{16} = \frac{8}{16} = \frac{1}{2}$$

$$P(X=2) = f(X=2,Y=0) + f(X=2,Y=1) + f(X=2,Y=2) = \frac{1}{16} + \frac{2}{16} + \frac{1}{16} = \frac{4}{16} = \frac{1}{4}$$

Another concept that is very important in regression analysis is the concept of statistically independent random variables. Two random variables X and Y are said to be statistically independent if and only if their joint probability mass function equals the product of their marginal probability functions for all combinations of X and Y:

$$f(X,Y) = f(X)f(Y) \text{ for all X and Y} \tag{1.8}$$

1.3 Characteristics of probability distributions

Even though the probability function for a random variable is informative and gives you all information you need about a random variable, it is sometime too much and too detailed. It is therefore convenient to summarize the distribution of the random variable by some basic statistics. Below we will shortly describe the most basic summary statistics for random variables and their probability distribution.

1.3.1 Measures of central tendency

There are several statistics that measure the central tendency of a distribution, but the single most important one is the expected value. The expected value of a discrete random variable is denoted E[X], and defined as follows:

$$E[X] = \sum_{i=1}^{n} x_i f(x_i) = \mu_X \tag{1.9}$$

It is interpreted as the mean, and refers to the mean of the population. It is simply a weighted average of all X-values that exist for the random variable where the corresponding probabilities work as weights.

Example 1.13

Use the marginal probability function in Example 1.12 and calculate the expected value of *X*.

$$E[X] = 0 \times P(X = 0) + 1 \times P(X = 1) + 2 \times P(X = 2) = 0.5 + 2 \times 0.25 = 1$$

When working with the expectation operator it is important to know some of its basic properties:

1) The expected value of a constant equals the constant, $E[c] = c$
2) If c is a constant and X is a random variable then: $E[cX] = cE[X]$
3) If a, b, and c are constants and X, and Y random variables then:
   $E[aX + bY + c] = aE[X] + bE[Y] + c$
4) If *X* and *Y* are statistically independent then and only then: $E[X, Y] = E[X]E[Y]$

The concept of expectation can easily be extended to the multivariate case. For the bivariate case we have

$$E[XY] = \sum_X \sum_Y XYf(X, Y) \tag{1.10}$$

Example 1.14

Calculate the $E[XY]$ using the information in Table 1.5. Following the formula we receive:

$$E[X,Y] = 0 \times 0 \times \frac{1}{16} + 0 \times 1 \times \frac{2}{16} + 0 \times 2 \times \frac{1}{16} + 1 \times 0 \times \frac{2}{16} + 1 \times 1 \times \frac{4}{16} + 1 \times 2 \times \frac{2}{16} +$$
$$2 \times 0 \times \frac{1}{16} + 2 \times 1 \times \frac{2}{16} + 2 \times 2 \times \frac{1}{16} = 1$$

1.3.2 Measures of dispersion

It is sometimes very important to know how much the random variable deviates from the expected value on average in the population. One measure that offers information about that is the variance and the corresponding standard deviation. The variance of *X* is defined as

$$Var[X] = \sigma_X^2 = E\left[(X - \mu_X)^2\right] = \sum_X (X - \mu_X)^2 f(X) \tag{1.11}$$

The positive square root of the variance is the standard deviation and represents the mean deviation from the expected value in the population. The most important properties of the variance is

1) The variance of a constant is zero. It has no variability.
2) If a and b are constants then $Var(aX + b) = Var(aX) = a^2 Var(X)$
3) Alternatively we have that Var(*X*) = E[X^2] – E[*X*]^2
4) E[X^2] = $\sum_x x^2 f(X)$

Example 1.15

Calculate the variance of X using the following probability distribution:

X	1	2	3	4
P(X)	1/10	2/10	3/10	4/10

Table 1.6 Probability distribution for X

In order to find the variance for X it is easiest to use the formula according to property 4 given above. We start by calculating E[X^2] and E[X].

$$\mathrm{E}[X] = 1\times\frac{1}{10}+2\times\frac{2}{10}+3\times\frac{3}{10}+4\times\frac{4}{10}=3$$

$$\mathrm{E}[X^2] = 1^2\times\frac{1}{10}+2^2\times\frac{2}{10}+3^2\times\frac{3}{10}+4^2\times\frac{4}{10}=10$$

$$\mathrm{Var}[X] = 10 - 3^2 = 1$$

1.3.3 Measures of linear relationship

A very important measure for a linear relationship between two random variables is the measure of covariance. The covariance of *X* and *Y* is defined as

$$Cov[X,Y] = E[(X - E[X])(Y - E(Y))] = E[XY] - E[X]E[Y] \tag{1.12}$$

The covariance is the measure of how much two random variables vary together. When two variables tend to vary in the same direction, that is, when the two variables tend to be above or below their expected value at the same time, we say that the covariance is positive. If they tend to vary in opposite direction, that is, when one tends to be above the expected value when the other is below its expected value, we have a negative covariance. If the covariance equals zero we say that there is no linear relationship between the two random variables.

Important properties of the covariance

1. $Cov[X,X] = Var[X]$
2. $Cov[X,Y] = Cov[Y,X]$
3. $Cov[cX\ ,Y] = cCov[X,Y]$
4. $Cov[X,Y+Z] = Cov[X,Y] + Cov[X,Z]$

The covariance measure is level dependent and has a range from minus infinity to plus infinity. That makes it very hard to compare two covariances between different pairs of variables. For that matter it is sometimes more convenient to standardize the covariance so that it become unit free and work within a much narrower range. One such standardization gives us the correlation between the two random variables.

The correlation between *X* and *Y* is defined as

$$Corr(X,Y) = \frac{Cov[X,Y]}{\sqrt{Var[X]Var[Y]}} \tag{1.13}$$

The correlation coefficient is a measure for the strength of the linear relationship and range from -1 to 1.

Example 1.16

Calculate the covariance and correlation for *X* and *Y* using the information from the joint probability mass function given in Table 1.7.

		Y			
		1	2	3	P(*X*)
	1	0	0.1	0	0.1
X	2	0.3	0.2	0.1	0.6
	3	0	0.3	0	0.3
	P(*Y*)	0.3	0.6	0.1	1.0

Table 1.7 The joint probability mass function for *X* and *Y*

We will start with the covariance. Hence we have to find E[*X*,*Y*], E[*X*] and [*Y*]. We have

$$E[X] = 1\times 0.1 + 2\times 0.6 + 3\times 0.3 = 2.2$$

$$E[Y] = 1\times 0.3 + 2\times 0.6 + 3\times 0.1 = 1.8$$

$$E[XY] = 1\times 1\times 0 + 1\times 2\times 0.1 + 1\times 3\times 0 + 2\times 1\times 0.3 + 2\times 2\times 0.2 + 2\times 3\times 0.1 +$$
$$3\times 1\times 0 + 3\times 2\times 0.3 + 3\times 3\times 0 = 4$$

This gives $Cov[X,Y] = 4 - 2.2\times 1.8 = 0.04 > 0$

We will now calculate the correlation coefficient. For that we need V[*X*], V[*Y*].

$$E[X^2] = 1^2\times 0.1 + 2^2\times 0.6 + 3^2\times 0.3 = 5.2$$

$$E[Y^2] = 1^2\times 0.3 + 2^2\times 0.6 + 3^2\times 0.1 = 3.6$$

$$V[X] = E[X^2] - E[X]^2 = 5.2 - 2.2^2 = 0.36$$

$$V[Y] = E[Y^2] - E[Y]^2 = 3.6 - 1.8^2 = 0.36$$

Using these calculations we may finally calculate the correlation using (1.13)

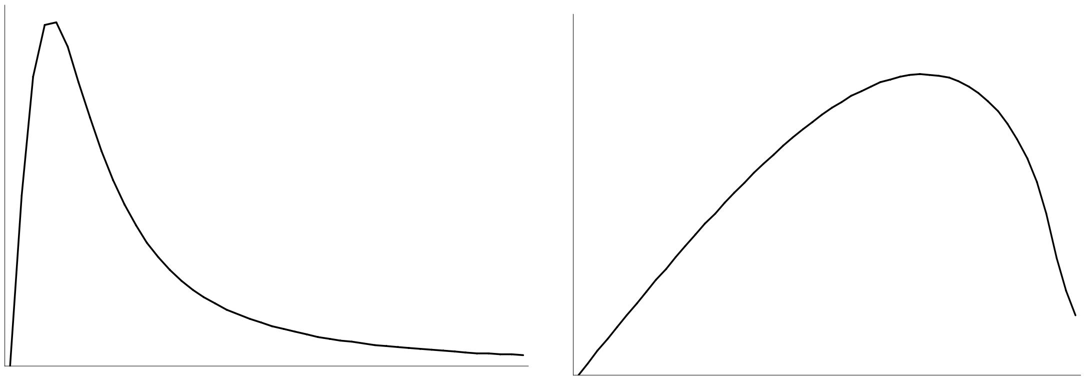
$$Corr[X,Y] = \frac{Cov[X,Y]}{\sqrt{V[X]V[Y]}} = \frac{0.04}{\sqrt{0.36\times 0.36}} = 0.11$$

1.3.4 Skewness and kurtosis

The last concepts that will be discussed in this chapter are related to the shape and the form of a probability distribution function. The Skewness of a distribution function is defined in the following way:

$$S = \frac{E[X-\mu_X]^3}{\sigma_X^3} \tag{1.14}$$

A distribution can be skewed to the left or to the right. If it is not skewed we say that the distribution is symmetric. Figure 1.1 give two examples for a continuous distribution function.



a) Skewed to the right b) Skewed to the left

Figure 1.1 Skewness of a continuous distribution

Kurtosis is a measure of whether the data are peaked or flat relative to a normal distribution. Formally it is defined in the following way:

$$K = \frac{E\left[X - \mu_X\right]^4}{\left[E\left[X - \mu_X\right]^2\right]^2} \tag{1.15}$$

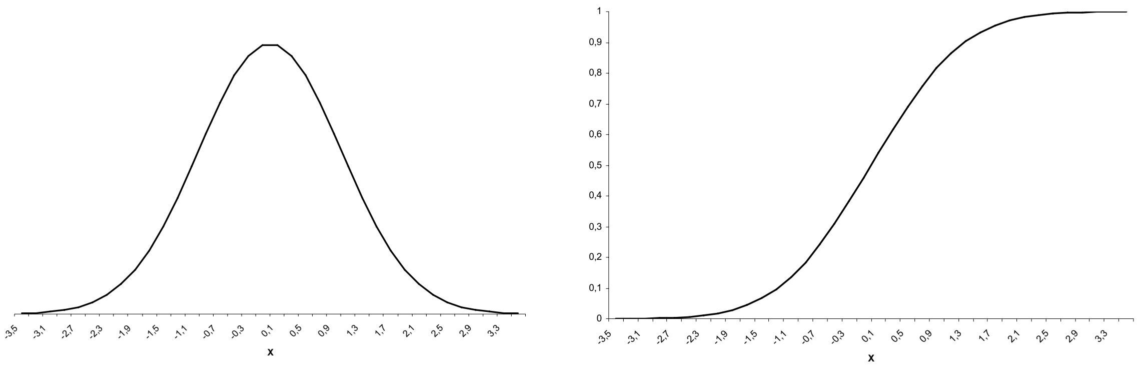
When a symmetric distribution follows the standard normal it has a kurtosis equal to 3. A distribution that are long tailed compared with the standard normal distribution has a kurtosis greater than 3 and if it is short tailed compared to the standard normal distribution it has a kurtosis that is less than three. It should be observed that many statistical programs standardize the kurtosis and presents the kurtosis as *K*-3 which means that a standard normal distribution receives a kurtosis of 0.

2 Basic probability distributions in econometrics

In the previous chapter we study the basics of probability distributions and how to use them when calculating probabilities. There exist a number of different probability distributions for discrete and continuous random variables, but some are more commonly used than others. In regression analysis and analysis related to regression analysis we primarily work with continuous probability distributions. For that matter we need to know something about the most basic probability functions related to continuous random variables. In this chapter we are going to work with the normal distribution, student t-distribution, the Chi-square distribution and the F-distribution function. Having knowledge about their properties we will be able to construct most of the tests required to make statistical inference using regression analysis.

2.1 The normal distribution

The single most important probability function for a continuous random variable in statistics and econometrics is the so called normal distribution function. It is a symmetric and bell shaped distribution function. Its Probability Density Function (PDF) and the corresponding Cumulative Distribution Function (CDF) are pictured in Figure 2.1.



a) Normal Probability Density Function

b) Normal Cumulative Distribution Function

Figure 2.1 The normal PDF and CDF

For notational convenience, we express a normally distributed random variable *X* as $X \sim N(\mu_X, \sigma_X^2)$, which says that *X* is normally distributed with the expected value given by μ_X and the variance given by σ_X^2. The mathematical expression for the normal density function is given by:

$$f(X) = \frac{1}{\sigma_X \sqrt{2\pi}} \exp\left\{ -\frac{1}{2}\left(\frac{X - \mu_X}{\sigma_X} \right)^2 \right\}$$

which should be used in order to determine the corresponding CDF:

$$P(X \leq c) = \int_{-\infty}^{c} f(X) dX$$

Unfortunately this integral has no closed form solution and need to be solved numerically. For that reason most basic textbooks in statistics and econometrics has statistical tables in their appendix giving the probability values for different values of c.

Properties of the normal distribution

1. The normal distribution curve is **symmetric** around its mean, μ_X, as shown in Figure 2.1a.
2. **Approximately 68%** of the area below the normal curve is covered by the interval of plus minus one standard deviation around its mean: $\mu_X \pm \sigma_X$.
3. **Approximately 95%** of the area below the normal curve is covered by the interval of plus minus two standard deviations around its mean: $\mu_X \pm 2 \times \sigma_X$.
4. **Approximately 99.7%** of the area below the normal curve is covered by the interval of plus minus three standard deviations around its mean: $\mu_X \pm 3 \times \sigma_X$.
5. A linear combination of two or more normal random variables is also normal.

Example 2.1

If X and Y are normally distributed variables, then $Z = aX + bY$ will also be a normally distributed random variable, where a and b are constants.

6. The **skewness** of a normal random variable is **zero**.
7. The **kurtosis** of a normal random variable equals **three**.
8. A **standard normal** random variable has a mean equal to zero and a standard deviation equal to one.
9. Any normal random variable X with mean μ_X and standard deviation σ_X can be transformed into a **standard normal** random variable Z using the formula $Z = \frac{X - \mu_X}{\sigma_X}$.

Example 2.2

Assume a random variable X with expected value equal to 4 and a standard deviation equal to 8. Using this information we may transform X into a standard normal random variable using the following transformation: $Z = \frac{X-4}{8}$. It is now easy to show that Z has a mean equal to 0 and a variance equal to 1. That is, we have

$$E[Z] = E\left[\frac{X-4}{8}\right] = E\left[\frac{X}{8}\right] - \frac{4}{8} = \frac{1}{8}E[X] - \frac{4}{8} = 0,$$

$$V[Z] = V\left[\frac{X-4}{8}\right] = V\left[\frac{X}{8}\right] = \frac{1}{64}V[X] = 1$$

Since any normally distributed random variable can be transformed into a standard normal random variable we do not need an infinite number of tables for all combinations of means and variances, but just one table that corresponds to the standard normal random variable.

Example 2.3

Assume that you have a normal random variable X with mean 4 and variance 9. Find the probability that X is less than 3.5. In order to solve this problem we first need to transform our normal random variable into a standard normal random variable, and thereafter use the table in the appendix to solve the problem. That is:

$$P(X \leq 3.5) = P\left(Z \leq \frac{3.5-4}{3}\right) = P(Z \leq -0.167)$$

We have a negative Z value, and the table does only contain positive values. We therefore need to transform our problem so that it adapts to the table we have access to. In order to do that, we need to recognize that the standard normal distribution is symmetric around its zero mean and the area of the pdf equals 1. That implies that $P(Z \leq -0.167) = P(Z \geq 0.167)$ and that $P(Z \geq 0.167) = 1 - P(Z \leq 0.167)$. In the last expression we have something that we will be able to find in the table. Hence, the solution is:

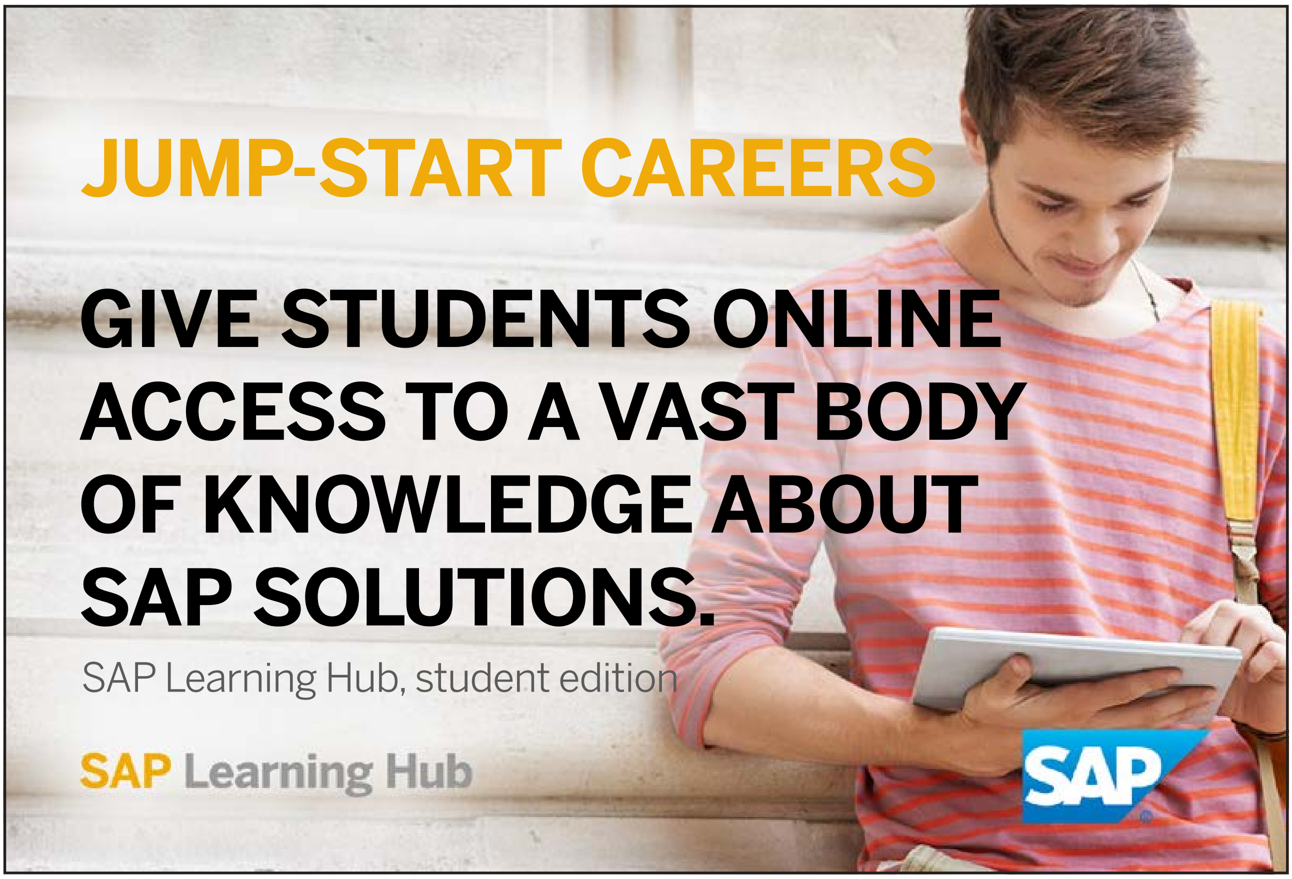
$$P(X \leq 3.5) = 1 - P(Z \leq 0.167) = 1 - 0.5675 = 0.4325$$

Example 2.4

Assume the same random variable as in the previous example and calculate the following probability: $P(3.5 \leq X \leq 4.5)$. Whenever dealing with intervals we need to split up the probability expression in two parts using the same logic as in the previous example. Hence, the probability may be rewritten in the following way:

$$P(3.5 \leq X \leq 4.5) = P(X \leq 4.5) - P(X \leq 3.5) = P\left(Z \leq \frac{4.5-4}{3}\right) - P\left(Z \leq \frac{3.5-4}{3}\right)$$
$$= P(Z \leq 0.167) - P(Z \leq -0.167)$$

In order to find the probability for this last equality we simply use the technique from the previous example.


JUMP-START CAREERS
GIVE STUDENTS ONLINE ACCESS TO A VAST BODY OF KNOWLEDGE ABOUT SAP SOLUTIONS.
SAP Learning Hub, student edition
SAP Learning Hub
SAP

The sampling distribution of the sample mean

Another very important concept in statistics and econometrics is the idea of a distribution of an estimator, such as the mean or the variance. It is essential when dealing with statistical inference. This issue will discussed substantially in later chapters and then in relation to estimators of the regression parameters.

The idea is quite simple. Whenever using a sample when estimating a population parameter we receive different estimate for each sample we use. That happens because of sampling variation. Since we are using different observations in each sample it is unlikely that the sample mean will be exactly the same for each sample taken. By calculating sample means from many different samples, we will be able to form a distribution of mean values. The question is whether it is possible to say something about this distribution without having to take a large number of samples and calculate their means. The answer is that we know something about that distribution.

In statistics we have a very important theorem that goes under the name **The Central Limit Theorem**. It says:

If *X1*, *X2*, … , *Xn* is a sufficiently large **random sample** from any population, with mean μ_X, and variance σ_X^2, then the distribution of sample means will be approximately normally distributed with $E[\overline{X}] = \mu_X$ and variance $V[\overline{X}] = \frac{\sigma_X^2}{n}$.

A basic rule of thumb says that if the sample is larger than 30 the shape of the distribution will be sufficiently close, and if the sample size is 100 or larger it will be more or less exactly normal. This basic theorem will be very helpful when carrying out tests related to sample means.

Basics steps in hypothesis testing

Assume that we would like to know if the sample mean of a random variable has changed from one year to another. In the first year we have population information about the mean and the variance. In the following year we would like to carry out a statistical test using a sample to see if the population mean has changed, as an alternative to collect the whole population yet another time. In order to carry out the statistical test we have to go through the following steps:

1. **Set up the hypothesis**
   In this step we have to form a null hypothesis that correspond to the situation of no change, and an alternative hypothesis, that correspond to a situation of a change. Formally we may write this in the following way:

   $$H_0 : \mu_X = \mu$$
   $$H_1 : \mu_X \neq \mu$$

In general we would like to express the hypothesis in such a way that we can reject the null hypothesis. If we do that we will be able to say something with a statistical certainty. If we are unable to reject the null hypothesis can just say that we do not have enough statistical material to say anything about the matter. The hypothesis given above is a so called a **two sided test**, since the alternative hypothesis is expressed with an inequality. The alternative would be to express the alternative hypothesis with larger than (>) or smaller than (<), which would resulted in a **one sided test**. In most cases, you should prefer to use a two sided test since it is more restrictive.

2. **Form the test function**

In this step we will use the ideas that come from the Central Limit Theorem. Since we have taken a sample and calculated a mean we know that the mean can be seen as a random variable that is normally distributed. Using this information we will be able to form the following test function:

$$Z = \frac{\overline{X} - \mu_X}{\sigma_X / \sqrt{n}} \sim N(0,1)$$

We transform the sample mean using the population information according to the null hypothesis. That will give us a new random variable, our test function *Z*, that is distributed according to the standard normal distribution. Observe that this is true only if our null hypothesis is true. We will discuss this issue further below.

3. **Choose the level of significance for the test and conclude**

At this point we have a random variable *Z*, and if the sample size is larger than 100, we know how it is distributed for certain. The fewer number of observations we have, the less we know about the distribution of Z, and the more likely it is to make a mistake when performing the test. In the following discussion we will assume that the sample size is sufficiently large so that the normal distribution is a good approximation.

Since we know the distribution of *Z*, we also know that realizations of *Z* take values between -1.96 and 1.96 in 95% of the cases (You should confirm this using Table A1 in the appendix). That is, if we take 100 samples and calculates the sample means and the corresponding test value for each sample, on average 95% of the test values will have values within this interval, **if our null hypothesis is correct**. This knowledge will now be used using only one sample.

If we take a sample and calculate a test value and find that the test value appear outside the interval, we say that this event is so unlikely to appear (less than 5 percent in the example above) that it cannot possible come from the distribution according to the null hypothesis (it cannot have the mean stated in the null hypothesis). We therefore say that we reject the null hypothesis in favor for the alternative hypothesis.

In this discussion we have chosen the interval [-1.96;1.96] which cover 95% of the probability distribution. We therefore say that we have chosen a 5% **significance level** for our test, and the end points for this interval are referred to as **critical values**. Alternatively, with a significance level of 5% there is a 5% chance that we will receive a value that is located outside the interval. Hence there is a 5% chance of making a mistake. If we believe this is a large probability, we may choose a lower significance level such as 1% or 0.1%. It is our choice as a test maker.

Example 2.5

Assume that you have taken a random sample of 10 observations from a normally distributed population and found that the sample mean equals 6. You happen to know that the population variance equals 2. You would like to know if the mean value of the population equals 5, or if it is different from 5.

You start by formulating the relevant null hypothesis and alternative hypothesis. For this example we have:

$$H_0 : \mu = 5$$
$$H_1 : \mu \neq 5$$

You know that according to the central limit theorem the sampling distribution of sample means has a normal distribution. We may therefore form the following test function:

$$Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} = \frac{6-5}{\sqrt{2/10}} = 2.236$$

We know that our test function follows the standard normal distribution (has a mean equal to zero) if the null hypothesis is true. Assume that we choose a significance level of 1%. A significance level of 1% means that there is a 1% chance that we will reject the null hypothesis even though the null hypothesis is correct. The critical values according to a significance level of 1% are [-2.576; 2.575].

Since our test value is located within this interval we cannot reject the null hypothesis. We have to conclude that the mean value of the population might be 5. We cannot say that it is significantly different from 5.

2.2 The t-distribution

The probability distribution that will be used most of the time in this book is the so called t-distribution. The t-distribution is very similar in shape to the normal distribution but works better for small samples. In large samples the t-distribution converges to the normal distribution.

Properties of the t-distribution

1. The t-distribution is symmetric around its mean.
2. The mean equals zero just as for the standard normal distribution.
3. The variance equals k/(k-2), with k being the degrees of freedom.

In the previous section we explained how we could transform a normal random variable with an arbitrary mean and an arbitrary variance into a standard normal variable. That was under condition that we knew the values of the population parameters. Often it is not possible to know the population variance, and we have to rely on the sample value. The transformation formula would then have a distribution that is different from the normal in small samples. It would instead be t-distributed.

Example 2.6

Assume that you have a sample of 60 observations and you found that the sample mean equals 5 and the sample variance equals 9. You would like to know if the population mean is different from 6. We state the following hypothesis:

$$H_0 : \mu = 6$$
$$H_1 : \mu \neq 6$$

We use the transformation formula to form the test function

$$t = \frac{\overline{X} - \mu_X}{S \big/ \sqrt{n}} \sim t_{(n-1)}$$

Observe that the expression for the standard deviation contains an *S*. *S* represents the sample standard deviation. Since it is based on a sample it is a random variable, just as the mean. The test function therefore contains two random variables. That implies more variation, and therefore a distribution that deviates from the standard normal. It is possible to show that the distribution of this test function follows the t-distribution with *n*-1 degrees of freedom, where *n* is the sample size. Hence in our case the test value equals

$$t = \frac{\overline{X} - \mu_X}{S \big/ \sqrt{n}} = \frac{5-6}{3 \big/ \sqrt{60}} = -2.58$$

The test value has to be compared with a critical value. If we choose a significance level of 5% the critical values according to the t-distribution would be [-2.0; 2.0]. Since the test value is located outside the interval we can say that we reject the null hypothesis in favor for the alternative hypothesis. That we have no information about the population mean is of no problem, because we assume that the population mean takes a value according to the null hypothesis. Hence, we assume that we know the true population mean. That is part of the test procedure.

2.3 The Chi-square distribution

Until now we have talked about the population mean and performed tests related to the mean. Often it is interesting to make inference about the population variance as well. For that purpose we are going to work with another distribution, the Chi-square distribution.

Statistical theory shows that the square root of a standard normal variable is distributed according to the Chi-square distribution and it is denoted χ^2, and has one degree of freedom. It turns out that the sum of squared independent standard normal variables also is Chi-squared distributed. We have:

$$Z_1^2 + Z_2^2 + \ldots + Z_k^2 \sim \chi^2_{(k)}$$

Properties of the Chi-squared distribution

1. The Chi-square distribution takes only positive values
2. It is skewed to the right in small samples, and converges to the normal distribution as the degrees of freedom goes to infinity
3. The mean value equals *k* and the variance equals 2*k*, where *k* is the degrees of freedom

In order to perform a test related to the variance of a population using the sample variance we need a test function with a known distribution that incorporates those components. In this case we may rely on statistical theory that shows that the following function would work:

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{(n-1)}$$

where S^2 represents the sample variance, σ^2 the population variance, and n-1 the degrees of freedom used to calculate the sample variance. How could this function be used to perform a test related to the population variance?

Example 2.7
We have a sample taken from a population where the population variance a given year was $\sigma^2 = 400$. Some years later we suspect that the population variance has increase and would like test if that is the case. We collect a sample of 25 observations and state the following hypothesis:

$$H_0 : \sigma^2 = 400$$

$$H_1 : \sigma^2 > 400$$

Using the 25 observations we found a sample variance equal to 600. Using this information we set up the test function and calculate the test value:

$$\text{Test Function } = \frac{(n-1)S^2}{\sigma^2} = \frac{(25-1)\times 600}{400} = 36$$

We decided to have a significance level of 5% and found a critical value in Table A3 equal to 36.415. Since the test value is lower than the critical value we cannot reject the null hypothesis. Hence we cannot say that the population variance has changed.

2.4 The F-distribution

The final distribution to be discussed in this chapter is the F-distribution. In shape it is very similar to the Chi-square distribution, but is a construction of a ratio of two independent Chi-squared distributed random variables. An F-distributed random variable therefore has two sets of degrees of freedom, since each variable in this ratio has its own degrees of freedom. That is:

$$\frac{\chi_m^2}{\chi_l^2} \sim F_{m,l}$$

Properties of the F-distribution

1. The F-distribution is skewed to the right and takes only positive values
2. The F-distribution converges to the normal distribution when the degrees of freedom become large
3. The square of a t-distributed random variable with *k* degrees of freedom become F-distributed: $t_k^2 = F_{1,k}$

The F-distribution can be used to test population variances. It is especially interesting when we would like to know if the variances from two different populations differ from each other. Statistical theory says that the ratio of two sample variances forms an F-distributed random variable with $n_1 - 1$ and $n_2 - 1$ degrees of freedom:

$$\frac{S_1^2}{S_2^2} \sim F_{(n_1-1)(n_2-1)}$$

Example 2.8

Assume that we have two independent populations and we would like to know if their variances are different from each other. We therefore take two samples, one from each population, and form the following hypothesis:

$$H_0 : \sigma_1^2 = \sigma_2^2$$

$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

Using the two samples we calculate the sample variances, $S_1^2 = 8.38$and $S_2^2 = 13.14$with $n_1 = 26$ and $n_2 = 30$. Under the null hypothesis we know that the ratio of the two sample variances is F-distributed with 25 and 29 degrees of freedom. Hence we form the test function and calculate the test value:

$$\frac{S_1^2}{S_2^2} = \frac{8.38}{13.14} = 0.638$$

This test value has to be compared with a critical value. Assume that we choose a significance level of 5%. Using Table A4 in the appendix, we have to find a critical value for a two sided test. Since the area outside the interval should sum up to 5%, we must find the upper critical point that corresponds to 2.5%. If we look for that value in the table we find 2.154. We call this upper point $F_{0.025}$. In order to find the lover point we can use the following formula:

$$F_{0.975} = \frac{1}{F_{0.025}} = \frac{1}{2.154} = 0.464$$

We have therefore received the following interval: [0.464;2.154]. The test value lies within this interval, which means that we are unable to reject the null hypothesis. It is therefore quite possible that the two population variances are the same.

3 The simple regression model

It is now time to leave the single variable analysis and move on to the main issue of the book, namely regression analysis. When looking at a single variable we could describe its behavior by using any summary statistic described in the previous chapters. Most often that would lead to a mean and a variance. The mean value would be a description of the central tendency, and the variance or the standard deviation a measure of how the average observation deviates from the mean. Furthermore, the kurtosis and skewness would say something about the distributional shape around the mean. But we can say nothing about the factors that make single observations deviate from the mean.

Regression analysis is a tool that can helps us to explain in part why observations deviate from the mean using other variables. The initial discussion will be related to models that use one single explanatory factor or variable *X* that explains why observations from the variable *Y* deviate from its mean. A regression model with only one explanatory variable is sometimes called **the simple regression model**. A simple regression model is seldom used in practice because economic variables are seldom explained by just one variable. However, all the intuition that we can receive from the simple model can be used in the multiple regression case. It is therefore important to have a good understanding of the simple model before moving on to more complicated models.

3.1 The population regression model

In regression analysis, just as in the analysis with a single variable, we make the distinction between the sample and the population. Since it is inconvenient to collect data for the whole population, we usually depends our analysis on a sample. Using this sample, we try to make inference on the population, that is, we try to find the value of the parameters that correspond to the population. It is therefore important to understand the distinction between the population regression equation and the sample regression equation.

3.1.1 The economic model

The econometric model, as appose to models in statistics in general, is connected to an economic model that motivate and explains the rational for the possible relation between the variables included in the analysis. However, the economic model is only a logical description of what the researcher believes is true. In order to confirm that the made assumptions are in accordance with the reality, it is important to specify a statistical model, based on the formulation of the economic model, and statistically test the hypothesis that the economic model propose using empirical data. However, it is the economic model that allows us to interpret the parameters of the statistical model in economic terms. It is therefore very important to remember that all econometric work has to start from an economic model.

Let us start with a very simple example. Economic theory claims that there is a relationship between food consumption and disposable income. It is believed that the monthly disposable income of the household has a positive effect on the monthly food expenditures of the household. That means that if the household disposable income increases, the food expenditure will increase as well.

To make it more general we claim that this is true in general, which means that when the average disposable income increase in the population, the average food expenditure will increase. Since we talk about averages we may express the economic model in terms of an expectation:

$$E[Y \mid X_1] = B_0 + B_1 X_1 \tag{3.1}$$

The conditional expectation given by (3.1) is a so called regression function and we call it the **population regression line**. We have imposed the assumption that the relationship between Y and X_1 is linear. That assumption is made for simplicity only, and later on when we allow for more variables, we may test if this is a reasonable assumption, or if we need to adjust for it. The parameters of interest are B_0 and B_1. In this text we will use capital letters for population parameters, and small letters will denote sample estimates of the population parameters. B_0 will represent the average food expenditure by households when the disposable income is zero $(X_1 = 0)$ and is usually referred to as the **intercept** or just the constant. The regression function also shows that if B_1 is different from zero and positive, the conditional mean of Y on X_1 will change and increase with the value of X_1. Furthermore, the **slope coefficient** will represent *the marginal propensity to spend on food*:

$$B_1 = \frac{dE[Y \mid X_1]}{dX_1}$$

3.1.2 The econometric model

We now have an economic model and we know how to interpret its parameters. It is therefore time to formulate the econometric model so that we will be able to estimate the size of the population parameters and test the implied hypothesis. The economic model is linear so we will be able to use linear regression analysis.

The function expressed by (3.1) represents an average individual. Hence when we collect data, individuals will typically not fall on the regression line. We might have households with the same disposable income, but with different level of food expenditures. It might even be the case that not a single observation is located on the regression line. This is something that we have to deal with. For the observer it might appear that the single observations locate randomly around the regression line. In statistical analysis we therefore control for the individual deviation from the regression line by adding a stochastic term (U) to (3.1), still under the assumption that the average observation will fall on the line. The econometric model is therefore:

$$Y_i = B_0 + B_1 X_{1i} + U_i \tag{3.2}$$

The formulation of the econometric model will now be true for all households, but the estimated population parameters will refer to the average household that is considered in the economic model. That is explicitly denoted by the subscript *i*, that appear on *Y*, X_1 and *U* but not on the parameters. We call expression (3.2) the **population regression equation**.

Adding a stochastic term may seem arbitrary, but it is in fact very important and attached with a number of assumptions that are important to fulfill. In the literature the name for the stochastic term differ from book to book and are called error term, residual term, disturbance term etc. In this text we will call the stochastic term of the population model for error term and when talking about the sample model we will refer to it as the residual term.

One important rational for the error term already mentioned is to make the equality hold true in equation (3.2) for all observations. The reason why it does not hold true in the first place could be due to omitted variables. It is quite reasonable to believe that many other variables are important determinants of the household food expenditure, such as family size, age composition of the household, education etc. There might in fact be a large number of factors that completely determines the food expenditure and some of them might be family specific. To be general we may say that:

$$Y = f(X_1, X_2, \ldots, X_k)$$

with *k* explanatory factors that completely determine the value of the dependent variable *Y*, where disposable income is just one of them. Hence, having access to only one explanatory variable we may write the complete model in the following way for a given household:

$$Y = B_0 + B_1 X_1 + f(X_2, X_3, ..., X_k)$$

$$Y = B_0 + B_1 X_1 + U$$

Hence everything left unaccounted for will be summarized in the term *U*, which will make the equality hold true. This way of thinking of the error term is very useful. However, even if we have access to all relevant variables, there is still some randomness left since human behavior is not totally predictable or rational. It is seldom the ambition of the researcher to include everything that accounts but just the most relevant. As a rule of thumb one should try to **have a model that is as simple as possible**, and avoid including variables with a combined effect that is very small, since it will serve little purpose. The model should be a simplistic version of the reality. The ambition is never to approach the reality with the model, since that will make the model too complicated.

Sometimes it might be the case that you have received data that has been rounded off, which will make the observations for the variable less precise. Errors of measurement are therefore yet another source of randomness that the researcher sometimes has no control over. If these measurements errors are made randomly over the sample, it is of minor problem. But if the size of the error is correlated with the dependent variable you might be in trouble. In chapter 4, book 2 we will discuss this issue thoroughly.

3.1.3 The assumptions of the simple regression model

The assumptions made on the population regression equation and on the error term in particular is important for the properties of the estimated parameters. It is therefore important to have a sound understanding of what the assumptions are and why they are important. The assumptions that we will state below is given for a given observation, which means that no subscripts will be used. That is very important to remember! The assumptions must hold for each observation.

Assumption 1: $Y = B_0 + B_1 X_1 + U$

The relation between *Y* and *X* is linear and the value of *Y* is determined for each value of *X*. This assumption also impose that the model is complete in the sense that all relevant variables has been included in the model.

Assumption 2: $E[Y \mid X] = B_0 + B_1 X_1$

$E[U \mid X] = E[U] = 0$

The conditional expectation of the residual is zero. Furthermore, there must not be any relation between the residual term and the *X* variable, which is to say that they are uncorrelated. This means that the variables left unaccounted for in the residual should have no relationship with the variable *X* included in the model.

Assumption 3: $V[Y] = V[U] = \sigma^2$

The variance of the error term is homoscedastic, that is the variance is constant over different observations. Since *Y* and *U* only differ by a constant their variance must be the same.

Assumption 4: $Cov(U_i, U_j) = Cov(Y_i, Y_j) = 0 \quad i \neq j$

The covariance between any pairs of error terms are equal to zero. When we have access to a randomly drawn sample from a population this will be the case.

Assumption 5: *X* need to vary in the sample.

X can not be a constant within a given sample since we are interested in how variation in *X* affects variation in *Y*. Furthermore, it is a mathematical necessity that *X* takes at least two different values in the sample. However, we going to assume that *X* is fixed from sample to sample. That means that the expected value of *X* is the variable itself, and the variance of *X* must be zero when working with the regression model. But within a sample there need to be variation. This assumption is often imposed to make the mathematics easier to deal with in introductory texts, and fortunately it has no affect on the nice properties of the OLS estimators that will be discussed at the end of this chapter.

Assumption 6: *U* is normally distributed with a mean and variance.

This assumption is necessary in small samples. The assumption affects the distribution of the estimated parameters. In order to perform test we need to know their distribution. When the sample is larger then 100 the distribution of the estimated parameters converges to the normal distribution. For that reason this assumption is often treated as optional in different text books.

Remember that when we are dealing with a sample, the error term is not observable. That means it is impossible to calculate its mean and variance with certainty, which makes it important to impose assumptions. Furthermore, these assumptions need to hold true for each single observation, and hence using only one observation to compute a mean and a variance is impossible.

3.2 Estimation of population parameters

We have specified an economic model, and the corresponding population regression equation. It is now time to estimate the value of the population parameters. For that purpose we need a **sample regression equation**, expressed as this:

$$Y_i = b_0 + b_1 X_{1i} + e_i \tag{3.3}$$

The important difference between the population regression equation and the sample regression equation concerns the parameters and the error term. In the population regression equation the parameters are fixed constants. They do not change. In the sample regression equation the parameters are random variables with a distribution. Their mean values represent estimates of the population parameters, and their standard errors are used when performing statistical test. The error term is also an estimate and corresponds to the population error term. Sometimes it is convenient to have a separate name for the estimated error term in order to make the distinction. In this text we will call the estimated error term the **residual term**.

THE ANSWER TO
YOUR LEARNING NEEDS
GET QUALITY, FLEXIBLE, AND ECONOMICAL TRAINING WHEN AND WHERE IT'S NEEDED.
SAP Learning Hub
SAP

3.2.1 The method of ordinary least squares

There exist many methods to estimate the parameters of the population regression equation. The most common ones are the method of maximum likelihood, the method of moment and the method of Ordinary Least Squares (OLS). The last method is by all means the most popular method used in the literature and is therefore the basis for this text.

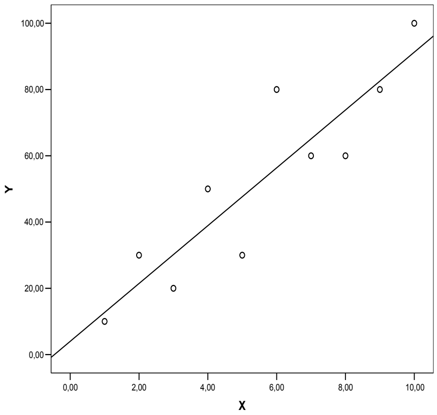


Figure 3.1 Fitted regression line using OLS

The OLS relies on the idea to select a line that represents an average relationship of the observed data similarly to the way the economic model is expressed. In Figure 3.1 we have a random sample of 10 observations. The OLS regression line is placed in such a way that the sum of the squared distances between the dots and the regression line become as small as possible. In mathematical terms using equation (3.3) we have:

$$RSS = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(Y_i - b_0 - b_1 X_{1i}\right)^2 \tag{3.4}$$

In order to be more general we assume a sample size of n observations. The objective is to minimize the Residual Sum of Squares (RSS) expressed in (3.4) with respect to b_0 and b_1. Hence, this is a standard optimization problem with two unknown variables that is solved by taking the partial derivatives with respect to b_0 and b_1, put them equal to zero, and then solving the resulting linear equations system with respect to those two variables. We have:

$$\frac{\partial RSS}{\partial b_0} = 2\sum_{i=1}^{n} \left(Y_i - b_0 - b_1 X_{1i}\right) = 0 \tag{3.5}$$

$$\frac{\partial RSS}{\partial b_1} = 2\sum_{i=1}^{n} \left(Y_i - b_0 - b_1 X_{1i}\right)\left(- X_{1i}\right) = 0 \tag{3.6}$$

By rearranging these two equations we obtain the equation system in normal form:

$$nb_0 + b_1\sum_{i=1}^{n} X_{1i} = \sum_{i=1}^{n} Y_i$$

$$b_0\sum_{i=1}^{n} X_i + b_1\sum_{i=1}^{n} X_{1i}^2 = \sum_{i=1}^{n} X_{1i}Y_i$$

Solving for b_0 and b_1 gives us:

$$b_0 = \overline{Y} - b_1\overline{X}_1 \tag{3.7}$$

$$b_1 = \frac{\sum_{i=1}^{n} X_{1i}Y_i - n\overline{X}\overline{Y}}{\sum_{i=1}^{n} X_{1i}^2 - n\overline{X}^2} = \frac{\sum_{i=1}^{n}\left(X_{1i} - \overline{X}_1\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{n}\left(X_{i1} - \overline{X}_1\right)^2} = \frac{Cov\left(X_1, Y\right)}{Var(X_1)} \tag{3.8}$$

The slope coefficient b_1 is simply a standardized covariance, with respect to the variation in X_1.The interpretation of this ratio is simply: when X_1 increases by 1 unit, Y will change by b_1 units. Remember that b_0 and b_1 are random variables, and hence it is important to know how their expected values and variances look like. Below we will derive the expected value and variance for both the intercept and the variance. The variance of the intercept is slightly more involved, but since text books in general avoid showing how it could be done we will do it here, even though the slope coefficient is the estimate of primary interest.

In order to find the expected value and the variance it is convenient to rewrite the expression for the estimators in such a way that they appear to be functions of the sample values of the dependent variable Y. Since the intercept is expressed as a function of the slope coefficient we will start with the slope estimator:

$$b_1 = \frac{\sum_{i=1}^{n}\left(X_{1i} - \overline{X}_1\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{n}\left(X_{i1} - \overline{X}_1\right)^2} = \frac{\sum_{i=1}^{n}\left(X_{1i} - \overline{X}_1\right)Y_i}{\sum_{i=1}^{n}\left(X_{i1} - \overline{X}_1\right)^2} = \sum_{i=1}^{n}\underbrace{\left[\frac{\left(X_{1i} - \overline{X}_1\right)}{\sum_{i=1}^{n}\left(X_{i1} - \overline{X}_1\right)^2}\right]}_{W_i}Y_i = \sum_{i=1}^{n} W_iY_i$$

$$b_1 = \sum_{i=1}^{n} W_iY_i \tag{3.9}$$

For the intercept we do the following:

$$b_0 = \overline{Y} - b_1\overline{X}_1 = \frac{1}{n}\sum_{i=1}^{n} Y_i - \overline{X}_1\underbrace{\sum_{i=1}^{n} W_iY_i}_{(3.9)} = \sum_{i=1}^{n}\left(\frac{1}{n} - \overline{X}_1W_i\right)Y_i = \sum_{i=1}^{n}\left(\frac{1}{n} - \overline{X}_1W_i\right)\left(B_0 + B_1X_{1i} + U_i\right) =$$

$$= B_0\underbrace{\sum_{i=1}^{n}\left(\frac{1}{n} - \overline{X}_1W_i\right)}_{=1} + B_1\underbrace{\sum_{i=1}^{n}\left(\frac{1}{n} - \overline{X}_1W_i\right)X_{1i}}_{=0} + \sum_{i=1}^{n}\left(\frac{1}{n} - \overline{X}_1W_i\right)U_i$$

Hence

$$b_0 = B_0 + \sum_{i=1}^{n}\left(\frac{1}{n} - \bar{X}_1 W_i\right)U_i \qquad (3.10)$$

Hence, the OLS estimators are weighted averages of the dependent variable, holding in mind that W_i is to be treated as a constant. Having the OLS estimators in this form we can easily find the expected value and variance:

The expected value of the OLS estimators

$$E[b_0] = B_0 + \sum_{i=1}^{n}\left(\frac{1}{n} - \bar{X}_1 W_i\right)E[U_i] = B_0 \qquad (3.11)$$

$$E[b_1] = E\left[\sum_{i=1}^{n} W_i Y_i\right] = E\left[\sum_{i=1}^{n} W_i\left(B_0 + B_1 X_{1i} + U_i\right)\right] = E\left[B_0 \underbrace{\sum_{i=1}^{n} W_i}_{=0}\right] + E\left[B_1 \underbrace{\sum_{i=1}^{n} W_i X_{1i}}_{=1}\right] + E\left[\sum_{i=1}^{n} W_i U_i\right]$$

and

$$E[b_1] = B_1 + \sum_{i=1}^{n} W_i \underbrace{E[U_i]}_{=0} = B_1 \qquad (3.12)$$

Hence, the mean value of the sample estimators equals the population parameters. You should confirm these steps by your self. The result from the second line comes from the regression assumptions. Also remember that the population parameter is a constant and that the expected value of a constant is the constant itself. The derivation of the variance will start with the expression established at the second line above.

The variance of the OLS estimators

When deriving the variance for the intercept, we utilize the definition of the variance that is expressed in terms of expectations. We have the expected value of the squared difference, and thereafter substitute

$$V\left[b_0\right] = E\left[b_0 - E(b_0)\right]^2 = E\left[\sum_{i=1}^{n}\left(\frac{1}{n} - \bar{X}_1 W_i\right)U_i\right]^2$$

Square the expression and take the expectation and end up with

$$V\left[b_0\right] = \left(\frac{1}{n} + \frac{\bar{X}_1^2}{\sum_{i=1}^{n}\left(X_{1i} - \bar{X}_1\right)^2}\right)\sigma^2 = \frac{\sigma^2\sum_{i=1}^{n}X_{1i}^2}{\sum_{i=1}^{n}\left(X_{i1} - \bar{X}_1\right)^2} \tag{3.13}$$

Try to work out the expressions and remember that $E\left[U_i^2\right] = E\left[U^2\right] = \sigma^2$ and that $E\left[U_iU_j\right] = Cov\left[U_i, U_j\right] = 0$.

$$V\left[b_1\right] = V\left[B_1 + \sum_{i=1}^{n}W_iU_i\right] = V\left[\sum_{i=1}^{n}W_iU_i\right] = \sum_{i=1}^{n}W_i^2V\left[U_i\right] = \sigma^2\sum_{i=1}^{n}\left[\frac{(X_{1i} - \bar{X}_1)}{\sum_{i=1}^{n}(X_{1i} - \bar{X}_1)^2}\right]^2 = \sigma^2\frac{\sum_{i=1}^{n}\left(X_{1i} - \bar{X}_1\right)^2}{\left[\sum_{i=1}^{n}(X_{1i} - \bar{X}_1)^2\right]^2}$$

and therefore

$$V\left[b_1\right] = \frac{\sigma^2}{\sum_{i=1}^{n}(X_{1i} - \bar{X}_1)^2} \tag{3.14}$$

The covariance between the two OLS estimators can be received using the covariance operator together with expressions (3.9) and (3.10). Try it out. The covariance is given by the following expression:

$$Cov(b_0, b_1) = \frac{-\bar{X}_1\sigma^2}{\sum_{i=1}^{n}\left(X_{1i} - \bar{X}_1\right)^2}$$

In order to understand all the steps made above you have to make sure you remember how the variance operator works. Go back to chapter 1, book 1 and repeat if necessary. Also remember that the variance of the population error term is constant and the same over observations. If that assumption is violated we will end up with something else.

Observe that the variance of the OLS estimators is a function of the variance of the error term of the model. The larger the variance of the error term, the larger becomes the variance of the OLS estimator. This is true for the variance of the intercept, variance of the slope coefficient and for the covariance between slope and the intercept. Remember that the variance of the error term and the variance of the dependent variable coincide. Also note that the larger the variation in X is, the smaller become the variance of the slope coefficient. Think about that. Increased variation in Y has of course the opposite effect, since the variance in Y is the same as the variance of the error term.

The variance of the population error term, σ^2, is usually unknown. We therefore need to replace it by an estimate, using sample information. Since the population error term is unobservable, one can use the estimated residual to find an estimate. We start by forming the residual term

$$e_i = Y_i - b_0 - b_1 X_{1i}$$

We observe that it takes two estimates to calculate its value which implies a loss of two degrees of freedom. With this information we may use the formula for the sample variance. That is:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$$

Observe that we have to divide by n-2, which referees to the degrees of freedom, which is the number of observations reduced with the number of estimated parameters used in order to create the residual. It turns out that this is an unbiased estimator of the population variance and it is decreasing as the number of observations increases.

3.2.2 Properties of the least squares estimator

The OLS estimator is attached to a number of good properties that is connected to the assumptions made on the regression model which is stated by a very important theorem; the **Gauss Markov theorem**.

The Gauss Markov Theorem

When the first 5 assumptions of the simple regression model are satisfied the parameter estimates are unbiased and have the smallest variance among other linear unbiased estimators. The estimators are then called **BLUE** for Best Linear Unbiased Estimators.

The OLS estimators will have the following properties when the assumptions of the regression function are fulfilled:

1) ***The estimators are unbiased***
   That the estimators are unbiased means that the expected value of the parameter equals the true population value. That means that if we take a number of samples and estimate the population parameters with these samples, the mean value of those estimates will equal the population value when the number of samples goes to infinity. Hence, on average we would be correct but it is not very likely that we will be exactly right for a given sample and a given set of parameters.

   Unbiased estimators implies that

$$\boxed{\begin{gathered} E[b_0] = B_0 \\ E[b_1] = B_1 \end{gathered}}$$

2) ***Minimum variance: Efficiency of unbiased estimators***
   When the variance is best, it means that it is efficient and that no other linear unbiased estimator has a better precision (smaller variance) of their estimators. It requires that the variance is homoscedastic and that it is not autocorrelated over time. Both these two issues will be discussed in chapter 1 and 2 in book 3.

3) ***Consistency***
   Consistency is another important property of the OLS estimator. It means that when the sample size increase and goes to infinity, the variance of the estimator has to converge to zero and the parameter converge to the population parameters. An estimator can be biased and still consistent but it is not possible for an estimator to be unbiased and inconsistent.

4) ***Normally distributed parameters***
   Since the parameters are weighted averages of the dependent variable they can be treated as a means. According to the central limit theorem, the distribution of means is normally distributed. Hence, the OLS estimators are normally distributed in sufficiently large samples.