Companion for "Statistics for Business and Economics" by Paul Newbold, William L. Carlson and Betty Thorne

Mynbaev, Kairat

Kazakh-British Technical University

4 June 2010

Companion

for "Statistics for Business and Economics"

by Paul Newbold, William L. Carlson and Betty Thorne

Kairat T. Mynbaev

International School of Economics

Kazakh-British Technical University

Tolebi 59, Almaty 050000, Kazakhstan

Chapter 0   Preface

Unit 0.1   What is the difference between the textbook and this companion?

I am going to call the book NCT, by the last initials of the authors. All references are to the sixth edition of NCT.

On the good side, the book has a lot of intuitive explanations and plenty of exercises, which is especially handy for instructors. The coverage of statistical topics is wide, so that most instructors will not need to look for other sources. Statistics requires a lot of math. The authors have done their best to hide the mathematical complexities and make the book accessible to those who are content using formulas without having to derive them.

This last advantage, looked at from a different angle, becomes a major drawback. If you have to remember just five formulas, after applying each once or twice perhaps you will memorize them. But how about 100 formulas? This is an approximate number of equations and verbal definitions (which eventually translate to equations) contained in the first six chapters of NCT. Some formulas are given more than once, in different forms and contexts. For example, different types of means are defined in Sections 3.1, 3.3, 5.3, 6.2 and 18.2, and it takes time and effort to relate them to one another. For your information, all types of means are special cases of just one, population mean. The origin of most equations is not explained. When most formulas fall out of the blue sky, after a while you get lost. When it is necessary to systematize testing procedures scattered over several sections or chapters, the situation becomes even more complex. Look at Figure 10.10 that describes the decision rule for choosing an appropriate testing procedure. If you are comfortable using it and all the relevant formulas without understanding the links between them, then you don't need this companion.

Now suppose that you are an ordinary human, who cannot mechanically memorize a bunch of formulas, like me or my students, who feel completely lost after the first six chapters if I follow the book's approach. Suppose you want to see a harmonious, coherent picture instead of a jigsaw puzzle of unrelated facts. Then I am going to try to convince you that studying formulas with proofs and derivations is the best option.

Unit 0.2   Why study math?

Firstly, it's not that much. You don't need to dig all the math you've forgotten from high school. Just study what I give (which is directly related to the course), and then you'll see that all properties of means make great sense and fit just two pages, with all the intuition, special cases and derivations. The same goes for variances, covariances, confidence intervals etc.

Secondly, knowing the logic (the links between theoretical facts) significantly reduces the burden on your memory. For example, just being familiar with the general principle of linearity will allow you to deduce all properties of covariances and variances in the logical chain

$$\text{means} \rightarrow \text{covariances} \rightarrow \text{variances}$$

from the properties of the first link (means). This is especially important in case of confidence intervals (and there are dozens of them) all of which are based on a few general ideas. My students love my handouts because instead of reading 30-40 pages in NCT they have to read just 9 pages (on average) for each chapter.

Thirdly, the knowledge cemented by logic stays in the memory much longer than mechanically memorized facts. Being able to derive equations you studied in the beginning of the semester will give you confidence when applying them during the final exam.

Fourthly, because in our brain everything is linked it has a wonderful property that can be called an *icebreaker principle*: as you develop one particular capability, many other brain activities enjoy a positive spillover effect. The logic and imagination you develop while studying math will give a comparative advantage in all your current and future activities that require analytical abilities. This influence on your skills and aptitude will be even more important than the knowledge of statistics as a subject.

Fifthly, the school approach to studying formulas through their application is not going to work with NCT. At high school you had to study a small number of different applications and had enough time to solve many exercises for each application type. In statistics, the number of different typical applications is so large that you will not have time to solve more than two exercises for each.

Sixthly, some students avoid math remembering their dreadful experience struggling with it at high school. Their fear of math continues well into college years, and they think: If I failed to learn math at high school, how can I learn it in just one semester (or year) at a university? I have a soothing and encouraging answer for such students. Most likely, you fell victim to a wrong teaching methodology, and I would not rely on your grades at high school to judge your math abilities. Some of your misunderstandings and problems with math may have come from your early age (when you saw for the first time, say, fractions or algebra rules). At that time your cognitive abilities were underdeveloped. Now you are an adult and, if you are reading this, you are a completely normal person with adequate cognitive abilities. You can comprehend everything much faster than when you were at elementary school. It's like with a human embryo, which in 9 months passes all the stages of evolution that took the humans ages to become who they are now. There are a couple of rules to follow though. Forget your fears, do not rely on that scanty information you are not sure about and look for common sense in everything you see. Relying on logic instead of memory is the best rule – this is why I prefer to teach people with work experience. Unlike fresh high-school graduates, they don't take anything for granted and are sincere about their lack of math background.

The **seventh** and the most important reason has been overlooked by all those who recommend to sidestep all formula derivation and concentrate on mastering formulas through their applications. See, some of us are endowed with good logic and imagination from birth. For such people, indeed, it is enough to see a formula and a model exercise for its application to be able to apply the formula on their own in a slightly different situation. But most of us mortals are not that bright. I've seen many people who understand well the statement of the exercise and know all the required theory and yet do not see the solution. In this case there is only one verdict: such individual's logic and imagination are not good enough. Explaining the solution to him/her usually doesn't help because that doesn't improve their logic and imagination and in a similar situation they will be lost again.

Logic and imagination are among the most advanced functions of our brain. Most people think that you either possess them or you don't and it is not possible to develop them. I can tell you a big secret that this is wrong, at least in math. Study mathematical facts and theories WITH the pertaining logic, and with time your own logic will start working (this is how your humble servant has become a mathematician). More about how to do this will be said later. For now let me state the main recommendation. Some of the exercises in NCT are pretty tricky and their diversity is amazing. Don't even hope to learn the solutions of a dozen model exercises and then be able to solve the rest of them.

> Your sure bet is to develop your own logic and imagination and the best way to achieve this purpose is to study mathematical facts with proofs and derivations.

OK, suppose I have convinced you that studying math is worthwhile and you are ready to give it a try. Then why not combine reading NCT with reading one of the extant mathematical texts? In what way is this manuscript different or better than the pile on the market? Well, this manuscript is not just another math text. It's the teaching philosophy that makes the difference. I talk about the learning strategy almost as much as about the subject itself. The exposition and study units are designed to develop your self-learning skills. There are comments on psychological aspects of the study process for students, teachers and even (present or future) parents.

Unit 0.3   Exposition

This manual is by no means a replacement for NCT. If I leave out some material, it means that it is well explained in NCT or it is not essential.

The order of definitions and statements in NCT does not reflect their logical sequence. Sometimes logically interrelated facts are scattered over different sections or even chapters in NCT. I do the opposite: the material is organized in study units, which combine logically linked information. This made necessary to give some material earlier or later than in NCT. Therefore placement of a study unit in a certain chapter is approximate. The heading of a unit contains the list of terms defined in it. Those headings and the index of terms allow you to establish the correspondence between NCT and this manual. I want my students to be able to answer questions of certain difficulty. Those questions determine the sizes of study units.

Unit 0.4   Formatting conventions

The level of difficulty of a unit is indicated by a line on the left side.

The dashed line on the left means something you are supposed to know from high school or math prerequisites.

There is no line on the left if the level roughly corresponds to NCT.

A single continuous line on the left means a higher level, which is desirable to study because it highlights important ideas. Sometimes it is included just to feed information-hungry students.

```
The text to type in Excel is shown in Courier font.
```

Italic indicates a new term even though the word "definition" may not be there.

> The most important methodological recommendations are framed like this.

Unit 0.5   Best ways to study

First best way. Study a theoretical problem, the existing approaches to its solution, try to improve or generalize upon them and write a publishable article. This way is not accessible to most students.

Second best way. As you study, write your own book or manual, in your own words and the way you understand it. For example, Joon Kang uploads to scribd.com his supplements to NCT with his own vision of the subject. This way is as rewarding as it is time-consuming (but quite accessible to some students). When I was 32, I wrote a book with my major professor M. Otelbaev. In one year I learned more than in the previous three years.

Third best way. After reading some material, close the book and write down the main points explaining the accompanying logic. The luckiest of you may have a classmate to retell the material to (I am a stickler of team-based learning). This way is quite feasible for most students. I call it *active recalling*, as opposed to *passive repetition*, when after reading the material you just skim it to see what you remember and what you don't. This way may be time-consuming initially but after a while you'll get used and your speed will increase. Many things you did not understand before will become simple. Try to work in the evenings when you are so tired that your memory refuses to function but your logic still works.

In the end the 1980's I had to give an intensive course in functional analysis. I gave 10 hours of lectures a week, 6 hours in a row in one day and 4 hours in another. As this all was high level stuff, the course was very useful for me (and a disaster for students).

Secret to success. Actively recall the material, in increasingly large chunks (not necessarily limited to my study units) and at an increasingly high speed.

If you don't do this, I don't guarantee anything. In my classes, I make sure that students actively repeat the material by arranging competitions among student teams.

Explanation. Nobody knows exactly what creative imagination means, in terms of neural activity. Let's talk about internal vision by which I mean the ability to see a large and complex picture at once. Internal vision can be compared to a ray of a flashlight. If the ray of your flashlight is narrow and the picture is big, you can see only a part of it. The wider the ray, the more of the picture you see. When you see the whole picture, you can solve the problem.

> Active recalling of increasingly large amounts of material forces the brain to adjust and gradually increase internal vision.

Increasing the speed of recalling (or writing) is important for two reasons. In addition to widening internal vision, it also strengthens the degree of excitation of neurons. Your thoughts become clearer and the speed of formation of new synapses (intra-neuron links) increases. The second reason is that high recalling speed reduces the degree of verbalization of the thinking process.

> In most cases, in math too much bla-bla-bla is harmful!
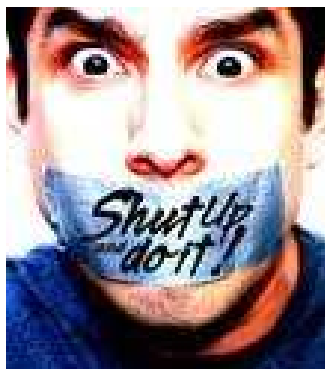
Send me feedback through LinkedIn.

Kairat Mynbaev

June 4, 2010

Chapter 1   How to Study Statistics?

Most of the material of the first two chapters of NCT can be safely skipped, not only because it is unimportant but primarily because our brain tries to systematize the incoming information, and there is very little systematic content in the first two chapters. Also, some issues from Chapters 1-2 are better taken up in subsequent chapters, when you are ready for their solution. I use this space to discuss some issues deeper than in the book and at the level you will need later. High-school material or the stuff you are supposed to remember from math prerequisites is reviewed. Solve the relevant exercises in the book only when you find my explanations insufficient.

Unit 1.1   The structure of mathematics (definitions, axioms, postulates, statements)

Math consists of definitions, axioms and statements. *Definitions* are simply names of objects. They don't require proofs. *Axioms* (also called *postulates*) are statements that we believe in; they don't require proofs and are in the very basement of a theory. *Statements* have to be proved. Depending on the situation and the tastes of the researcher, they may have different other names: theorem, lemma, proposition, criterion etc. Statements are a way to economize on space and thinking efforts. They summarize sometimes very long arguments and serve as building bricks in subsequent constructions.

In math texts definitions take up at most 5 to 10% of space. A common error of students of quantitative subjects is to jump to statements without thinking enough about definitions[1]. Definitions not only give names to objects but they also give direction to the theory and reflect the researcher's point of view. Often understanding definitions well allows you to guess or even prove some results.

Unit 1.2   Studying a definition (natural, even, odd, integer and real numbers; sets)

A definition starts with a preamble which sets a background and allows you to see the existence of objects with distinct properties. Most of the time the preamble is not even mentioned but there are cases when it contains a complex logical argument. After you understand the preamble understanding the definition proper – that is giving the name – becomes easy. Studying a definition is concluded by considering possible equivalent definitions and deriving immediate consequences.

Example 1.1 (1) Preamble. Let us consider natural numbers (these are numbers 1, 2, 3, … that naturally arise when counting things). We can notice that some of them are divisible by 2 (for example, $4/2 = 2$ is an integral number) and others are not (e.g., $3/2$ is a fractional number).

(2) Definition proper. The natural numbers that are divisible by 2 (2, 4, 6, …) are called *even*. The natural numbers that are not divisible by 2 (1, 3, 5, …) are called *odd*.

---

[1] When I was young I also committed this error.

(3) Immediate consequences. (i) Any natural number is either even or odd. (ii) If a natural number is even and divisible by 3 at the same time, then it is divisible by 6.

Some basic objects cannot be defined or their precise definitions are too complex for this course. In such cases it is better to give equivalent (descriptive) names or simply list the objects we name (as it has been done in case of natural numbers).

Example 1.2 The following names are used interchangeably: set = collection = family = array.

Example 1.3 *Natural numbers* are members of the set $N = \{1, 2, 3, ...\}$. *Integer numbers* are members of the set $Z = \{0, \pm 1, \pm 2, ...\}$. In some applications (see Unit 5.12 and Unit 5.15) we shall need nonnegative integer numbers listed in $Z_+ = \{0, 1, 2, ...\}$.

Example 1.4 The next type of numbers we need, real numbers, cannot be put in a list. As a convenient working definition, we call *real numbers* those numbers that can be written in a decimal form. For example, the famous number π, is, by definition, the ratio of the length of a circumference to its diameter. Its decimal form, is, up to approximation, $\pi = 3.1415926...$ Integer numbers fall into this category: for example, $1 = 1.000...$ The set of all real numbers is denoted R.

Unit 1.3   Ways to think about things (commutativity rules, quadratic equation, real line, coordinate plane, Venn diagrams)

(i) When you study a definition, don't think about a single representative of a class of objects; try to think about the whole class. Thinking in terms of sets will systematize your knowledge.

(ii) In algebra, when we denote numbers by letters, we mean that whatever rules we write with letters are of universal character and apply to all numbers in a given set. For example, the *commutativity rules* for summation and multiplication $a + b = b + a$, $ab = ba$ hold for all real numbers and not just for those specific ones you may plug in. Simple algebra rules, like the ones we have just seen, are easy to formulate verbally. Verbal formulation and derivation of complex algebra rules is a nightmare. For example, try to verbalize the rule for finding the roots of the *quadratic equation* $ax^2 + bx + c = 0$. This is what people did before 1637, when René Descartes introduced modern algebraic notation. This is what people still do in the beginning and intermediate micro- and macroeconomic courses. Excuse me, but this is a stone age! This tradition persists for the only reason that the teaching methodology at secondary and high school fails. With the right methodology, every normal person can learn algebra[2].

(iii) Believe me, even with zero math background you can quickly learn algebra if you follow a few simple rules:

> **(a)** Pay attention to small details, like the usage of upper- and lower-case letters, the difference between parentheses (), brackets [] and braces {}, punctuation and arithmetic signs (+, −, ×, /).

---

[2] I studied early education issues, experimented with my children and know what I am talking about.

(b) Try to find a tangible (or intuitive) interpretation for each element in a mathematical construction. Usually, geometry clarifies many things. For example, a real number can be uniquely identified with a point on a straight line (called a *real line* in this case); a pair of real numbers can be uniquely identified with a point on a *coordinate plane*; *Venn diagrams* help when working with sets.

(c) Similarly, explain for yourself each step in a long algebraic transformation. Don't trust the book or the professor.

> Your knowledge of and interest in the subject will directly depend on the percentage of the material you can explain.

(d) Initially it helps to translate formulas to verbal statements and back. Verbalization can also help in case of long, complex definitions and statements (we'll return to this when discussing p-values, see Unit 10.4). With time you'll be able to do without verbalization, and in complex situations the best thing to do is shut up and concentrate[3].

(e) In such a theoretical subject as statistics it is not possible to provide real-life motivating examples for absolutely all parts of the theory. In fact, many theoretical problems are prompted by the internal logic of the theory; even more often the solutions to the problems are based on mathematical intuition. We are not computers; we need some level of interest in the subject and some level of satisfaction from our efforts to continue our studies. It may be relieving to learn that

> We tend to like what we understand.

Try to keep your level of understanding high (say, at 80 to 90%) and the interest in the subject will ensue.

(f) If your ambition is to get a high mark in statistics, follow what I call a *110% rule*. We inevitably forget a part of what we've studied.

> To know all 100% of the material, you have to study 110%.

---

[3] When I brought my son to the USA and put him to the 7th grade, he did not speak English. The math teacher would explain to the students the solution for half an hour, they would diligently write down his explanations and my son would sit doing nothing. The teacher would ask why he wasn't writing anything and my son would show a one-line solution to the problem.

Chapter 2   Describing Data: Graphical

Unit 2.1   Classification of variables (definitions: intuitive, formal, working, simplified, descriptive, complementary; variables: numerical, categorical, discrete, continuous)

An important part of working with definitions consists in comparing related definitions. This is the right time to discuss the choice from among different definitions of the same object. An *intuitive definition* gives you the main idea and often helps you come up with the *formal definition*, which is mostly a formula[4]. Formal definitions may have different forms. The one which is easier to apply is called a *working definition*. Some complex definitions may be difficult to understand on the first push. In this case it is better to start with *simplified definitions*. They usually are also shorter. We also distinguish between *descriptive definitions* (when properties of an object are described) and *complementary definitions* (instead of saying what an object is we can say what it is not).

In the next diagram you see a summary of Section 2.1 from NCT on classification of variables. Those definitions that are not essential for the subsequent chapters are skipped. For the remaining I give short, working and simplified definitions.

```
                         ┌──────────────┐
                         │  Variables   │
                         └──────┬───────┘
            ┌───────────────────┴───────────────────┐
  ┌─────────────────────────┐         ┌─────────────────────────┐
  │ Numerical (the ones     │         │ Categorical (whose       │
  │ which take numerical     │         │ values are not numerical)│
  │ values)                  │         │                          │
  └──────────┬──────────────┘         └──────────────────────────┘
     ┌───────┴────────┐
┌──────────────────┐  ┌──────────────────┐
│ Discrete         │  │ Continuous       │
│ (simplification: │  │ (which take all  │
│ the ones that    │  │ values in some   │
│ take a finite    │  │ interval of a    │
│ number of values)│  │ real line)       │
└──────────────────┘  └──────────────────┘
```
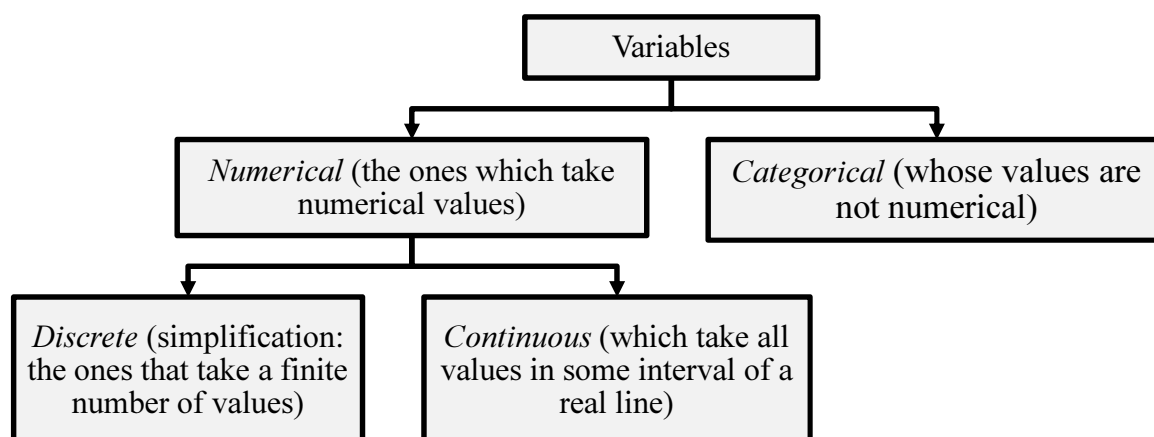
NCT usually give several motivating examples. I recommend remembering and thinking about just one motivating example for each definition (or statement, or method).

Problem with categorical variables. Categorical variables have been defined as complementary to numerical ones. Most statistical measures use arithmetic operations. Therefore they are not applicable to categorical variables. The workaround is to associate numerical variables with categorical variables. For example, instead of working with responses "Yes" and "No" we can work with the variable that takes two values: 1 for "Yes" and 0 for "No".

---

[4] The mathematician Kronecker is reported to have said," God created the natural numbers; all the rest is the work of man." Read more: Natural Numbers: http://science.jrank.org/pages/4570/Natural-Numbers.html#ixzz0lznifXd9.

Unit 2.2   Frequencies and distributions (Bernoulli variable, binomial variable, sample size, random or stochastic variable, absolute and relative frequencies, frequency distributions)

One of the implicit purposes of Chapter 2 is to show you that real-life data are huge and impossible to grasp without certain graphical tools or summary measures. I think at this point it is important for you to get an idea of how the data arrive and how they are transformed for the purposes of statistical data processing. Using an Excel spreadsheet will give you the feeling of the data generation process in real time[5].

Exercise 2.1 We are going to simulate a numerical two-valued variable (in Unit 7.3 it will be identified as the *Bernoulli variable*). In cell A1 type the name `Bernoulli`. In cell A2 enter the formula

```
=IF(RAND()>1/4,0,1)
```

exactly as it is (no blanks)[6]. You can use lower-case letters but Excel will convert them to upper-case. In Tools/Options/Calculations select "Manual" and click OK. Now press F9 several times to see some realized values of the Bernoulli variable.

Exercise 2.2 This example models a variable with 4 integer values (we'll see in Unit 7.3 that it's a *binomial variable*). Copy the formula you entered in A2 to cells A3, A4 (you can select cell A2, pick with a mouse a small square in the lower right corner of the cell and pull it down for the selection to include cells A3, A4; the formula will be copied). In cell B1 type `Binomial` and in cell B2 enter the formula

```
=SUM(A2:A4)
```

By pressing F9 you can observe the realized values of the binomial variable. Many students consider it quite a headache. It will be, if you jump to the conclusions.

Exercise 2.3 This is the main exercise. We are going to collect a sample of observations on the binomial variable. By definition, the *sample size*, n, is the number of observations.

(a) Choose n, press F9 n times and write down the values observed in cell B2:

$$b_1 = ..., \ b_2 = ..., ..., b_n = ... \tag{2.1}$$

Don't be lazy and use at least $n = 30$.

(b) After having pressed F9 n times do you think you can predict what will show up next time? **Intuitive definition** of a random variable: a variable is called *random* (or *stochastic*) if its values cannot be predicted.

(c) Alternatively, do you think you can tell what is NOT likely to show up next time? The answer I am pushing for is that any number except 0, 1, 2, 3 is not likely to appear.

---

[5] I am using Office 2003; all simulations are done in the same file; for new examples I use the cells to the right of the filled ones.

[6] The explanation of all Excel functions we use can be found in Excel Help. The mathematical side of what we do will be explained in Unit 7.3.

(d) Denote x_i distinct values in your sample. Most probably, they are $x_1 = 0, x_2 = 1, x_3 = 2, x_3 = 3$. Count the number of times x_i appears in your sample and denote it n_i, $i = 1, 2, 3, 4$. These numbers are called *absolute frequencies*. Do you think their total is n:

$$n_1 + n_2 + n_3 + n_4 = n? \tag{2.2}$$

(e) The numbers $r_i = n_i / n$, $i = 1, 2, 3, 4$, are called *relative frequencies.* r_1, for example, shows the percentage of times the value x_1 appears in your sample. Do you think the relative frequencies sum to 1:

$$r_1 + r_2 + r_3 + r_4 = 1? \tag{2.3}$$

You can get this equation from (2.2) by dividing both sides by n.

(f) Summarize your findings in a table:

Table 2.1 Table of absolute and relative frequencies (discrete variable)

Values of the variable	Absolute frequencies	Relative frequencies
$x_1 = 0$	$n_1 = \ldots$	$r_1 = \ldots$
$x_2 = 1$	$n_2 = \ldots$	$r_2 = \ldots$
$x_3 = 2$	$n_3 = \ldots$	$r_3 = \ldots$
$x_4 = 0$	$n_4 = \ldots$	$r_4 = \ldots$
	Total = n	Total = 1

Conclusions. Suppose somebody else works with a different sample size. Your and that person's absolute frequencies will not be comparable because their totals are different. On the other hand, comparing relative frequencies will make sense because they sum to 1. Relative frequencies are better also because they are percentages. The columns of frequencies show how totals are distributed over the values of the variable. Therefore they are called *absolute* and *relative frequency distributions*, respectively. For convenience reasons, most of the time the last column of **Table 2.1** is used and the adjective "relative" is omitted. Once you understand this basic situation, it is easy to understand possible variations, which we consider next.

Categorical variables. For example, suppose that at a given university there are four instructor ranks: lecturer, assistant professor, associate professor and full professor. Then instead of **Table 2.1** we have

Table 2.2 Table of absolute and relative frequencies (categorical variable)

Values of the variable	Absolute frequencies	Relative frequencies
Lecturer	$n_1 = \ldots$	$r_1 = \ldots$
Assistant professor	$n_2 = \ldots$	$r_2 = \ldots$
Associate professor	$n_3 = \ldots$	$r_3 = \ldots$
Full professor	$n_4 = \ldots$	$r_4 = \ldots$
	Total = n	Total = 1

Exercise 2.4 Here we model a continuous variable. See Unit 7.3 for the theory.

(a) To model summer temperatures in my city, type in cell C1 `Temperature` and in cell C2 the formula

16

```
=NORMINV(RAND(),25,8)
```

The resulting numbers are real and may have many nonzero digits. Usually nobody reports temperatures with high precision. In Format/Cells/Number select "Number" as the category and 1 decimal place in "Decimal places" and click OK.

(b) Select the sample size and write down the observed values in form (2.1). Since the elements in the sample are real numbers, none may be repeated and in that case all absolute frequencies are equal to 1. What is worse, if you obtain another sample, it may have very few common elements with the first one. To achieve stability over samples and reduce the table size, in such cases the observations are joined in ranges (groups, clusters). In case of temperatures you can use intervals of length 1:

$$t_{min} = 10 \le t < 11, \ ..., \ 39 \le t < 40 = t_{max} \tag{2.4}$$

(the values t_{min}, t_{max} will in fact depend on the sample and may be different from the ones you see here). In this approach the values in the first column will be intervals (2.4).

(c) For each observation, you calculate the number of observations falling into that interval to find the absolute frequency. Since observed values are replaced by the intervals they fall into, under this approach a part of the information contained in the sample is lost. However, if the lengths of the intervals are decreased, the loss decreases and the resulting frequencies represent the variables very well. Denoting k the number of intervals, we obtain a table of type:

Table 2.3 Table of absolute and relative frequencies (continuous variable)

Values of the variable	Absolute frequencies	Relative frequencies
$10 \le t < 11$	$n_1 = ...$	$r_1 = ...$
…	…	…
$39 \le t < 40$	$n_k = ...$	$r_k = ...$
	Total = n	Total = 1

Unit 2.3 Visualizing statistical data (coordinate plane; argument, values, domain and range of a function; independent and dependent variables; histogram, Pareto diagram, time series, time series plot, stem-and-leaf display, scatterplot)

Some graph types used by statisticians are readily available in Excel. Others require some data manipulation to produce results equivalent to those available in Minitab. Alternatively, you can buy SPC XL, a statistical add-in for Excel. But I think discussing definitions of main graph types is more important than seeing them.

In the coordinate plane, we have two axes. x-values are put on the horizontal axis. y-values are put on the vertical axis. A point with coordinates (x, y) is located at the intersection of two straight lines: one of them is drawn vertically through point $(x, 0)$ (which is on the x-axis); the other is drawn horizontally through point $(0, y)$ (which is on the y-axis).

A graph is the best tool to visualize a functional relationship. The rough-and-tough approach to graphing a function $y = f(x)$ is to (a) select several values of the argument $x_1, ..., x_n$, (b) calculate the corresponding values of the function $y_1 = f(x_1), ..., y_n = f(x_n)$, (c) put the

points (x_1, y_1), ..., (x_n, y_n) on the coordinate plane and (d) draw a smooth line through these points. If you have never done this, do it at least once, say, for $y = 2x^2 + 1$.

Keep in mind the terminology introduced here: x is the *argument*, $f(x)$ is the *value*. The set of all arguments of the given function is called its *domain*. The set of all values is called its *range*. As the argument runs over the domain, the value runs over the range.

> When working with functions, it's better to use words that emphasize motion. The notion of a function is one of the channels through which motion is introduced in mathematics.

The argument is also called an independent variable, while the value – a dependent variable. It is customary to plot the dependent variable on the y-axis and the independent one on the x-axis, the most notable exceptions being the plots of demand and supply in economics.

In order to be at sea with statistical graphs, you have to be clear about what is the argument and what is the value in each case. The short way to express this is: plot this … (values) against that … (arguments).

2.3.1 *Histogram* short definition: plot frequencies against values.

Explanations. **(i)** In case of numerical values, put them on the x-axis. Atop each of them show the corresponding frequency with a vertical bar whose height is equal to the frequency. The width of the bar doesn't matter, as long as the bars have the same widths and don't overlap. You can use either absolute or relative frequencies. I prefer relative frequencies and from now on talk only about them.

(ii) When values are intervals, we know from **Exercise 2.4** that intervals of type (2.4) are adjacent. Bars showing frequencies are plotted against centers of the intervals. Regarding the heights and widths of the bars there are two possibilities:

> (ii$_a$) The heights express frequencies. The widths don't matter as long as the bars don't overlap.

> (ii$_b$) The areas of the bars are equal to frequencies. In this case the widths are equal to the lengths of the intervals on the x axis (the bars stand shoulder to shoulder) and the heights are calculated from the usual area rule

$$\text{area} = \text{height} \times \text{width}.$$

The second possibility is used when, as in our case, the sample is obtained from a continuous variable. As we'll see in Unit 6.2, for a continuous variable there is an equivalent of relative frequencies, and that equivalent hinges upon area rather than height.

(iii) Suppose the values are categories other than ranges of numerical values, like instructor ranks. You put them on the x-axis at arbitrary points and plot the bars as in case (i).

2.3.2 *Pareto diagram*: same as a histogram, except that the observations are put in the order of decreasing frequencies.

Example 2.1 In this case it is important to remember the motivation. Suppose there are too many traffic jams in the city and the mayor is determined to eliminate them. Among the reasons of traffic jams there are: lack of parking space in the center of the city, car accidents, absence of car junctions in critical areas, rush hours etc. It makes sense to start dealing with problems that have the highest impact on traffic jams frequency. Such problems (categories) on the histogram should be put first.

2.3.3 *Time series plot*: plot values against time.

Exercise 2.5 A *time series* is a sequence of observations on a variable along time. For example, a sequence of daily observations on Microsoft stock can be written as $s_1, ..., s_n$. The observations are plotted in their raw form, without calculating any frequencies. We model them in Excel as follows. In D1 type `Stock` and in D2 enter the formula

$$=\text{NORMINV(RAND(),25,8)}$$

Copy it to cells D3-D21 (which corresponds to $n = 20$). Select column D and in the Chart Wizard select "Line" for the plot type. When you press F9, you see various realizations of the time series. Time series plots are good to see if there are any tendencies or trends in a series and how volatile it is. Financial time series are among the most volatile.

2.3.4 *Stem-and-leaf display*

Firstly, this is not a graphical display. It is a tabular representation of the data where the observations are grouped according to their leading digits and the final digits (remainders) are used to show frequencies in each group. Secondly, going from the usual decimal representation to the stem-and-leaf display and back is drudgery. Given its very limited use in statistical data processing and availability of computers, having students do it by hand is a stone age. The only reason I give the next exercise is that in NCT it is not explained how you go from the stem-and-leaf display back to the decimal representation.

Example 2.2 Let's say we have a sample of lengths of 10 fish:

$$x_1 = 9.8, \ x_2 = 10.1, \ x_3 = 14.5, \ x_4 = 10.5, \ x_5 = 13.2,$$
$$x_6 = 11.1, \ x_7 = 9.2, \ x_8 = 11.7, \ x_9 = 15.0, \ x_{10} = 10.2.$$

We take the integer part (the number before the decimal point) as the *stem* (leading digits) and the digit after the decimal point as the *leaf* (final digits). The lengths $x_2 = 10.1, \ x_4 = 10.5, \ x_{10} = 10.2$ fall into one group. The stems are shown in the second column and the leaves in the third column. The first column contains cumulative frequencies. To obtain them, you start counting the absolute frequencies from the upper and lower ends until the cumulative frequencies are about the same (in the middle). The cumulative frequency in the first row is the absolute frequency in that row; the cumulative frequency in the second row is the sum of absolute frequencies in that row and the preceding row. You go like that and stop at row 4 because if you count similarly the cumulative frequencies starting from the bottom, in row 5 you will have 8.

Table 2.4 Stem-and-leaf display

Cumulative frequencies	Stem	Leaf
2	9	2, 8
5	10	1, 5, 6
7	11	1, 7
7	12	
8	13	2
6	14	5
1	15	0
	Total = n	Total = 1

Note four features of this display.

(i) The stems are ordered, unlike the original data.

(ii) The number of leaves readily shows absolute frequencies in each group.

(iii) You can restore the original values from this display if you know how much one unit in the stem is worth. In our example it is worth $w=1$ so, for example, the first line gives two values

$$9 \cdot w + 2 \cdot w \cdot 0.1 = 9 + 0.2 = 9.2, \ \ 9 \cdot w + 8 \cdot w \cdot 0.1 = 9 + 0.2 = 9.8.$$

(iv) If we considered young fish, instead of 9.2 and 9.8 we could have 0.92 and 0.98 and each unit in the stem would have been worth $w=0.1$. The above calculation would have given the right values.

2.3.5 *Scatterplot*: plot values of one variable against values of another.

> Be alert when you see a discussion of pairs of variables. Most students find this topic difficult.

Exercise 2.6 We want to see a rough dependence between the weight and height of an adult. For a sample of n adults we write down their weights w_1, ..., w_n and heights h_1, ..., h_n. The first thing to note is that these observations should be written in pairs because it doesn't make sense to correlate one person's weight with another person's height. The second thing is that, logically, it is the height that determines the weight and not the other way around. Therefore the right way to write the observations is

$$(h_1, w_1), \ ..., \ (h_n, w_n). \tag{2.5}$$

A *scatterplot* just shows these points on the coordinate plane. Now we proceed with simulating them.

(i) In cell E1 type `Heights` and in cell E2 put the formula

```
=NORMINV(RAND(),170,30)
```
(heights are measured in centimeters and weights in kilos). In cell F1 type `Weights` and in cell F2 put the formula
```
=E2-100+NORMINV(RAND(),0,20)
```

Select cells E2, F2 and by pulling down the small square in the lower right corner of cell F2 copy the contents of E2, F2 to sufficiently many cells, including E21, F21 (for future exercises).

(ii) Select columns E, F and use the plot type X,Y (Scatter) without lines connecting the points to see the scatterplot. To make it nicer, in Step 3 of the Chart Wizard on the tab "Titles" type `Heights` for Value (X) axis and `Weights` for Value (Y) axis; on the tab "Legend" uncheck "Show legend".

(iii) You can see on the plot that there is, on average, a positive relationship between weight and height, as expected. If you press F9, the sample and plot will change.

Unit 2.4   Questions for repetition

1. Describe verbally how you get Table 2.1.

2. Give short definitions of a histogram, Pareto diagram, time series plot, stem-and-leaf display, scatterplot and indicate when they are appropriate.

Chapter 3   Describing Data: Numerical

Unit 3.1   Three representations of data: raw, ordered and frequency representation

In this chapter we consider numerical characteristics of data sets. As you think about the definitions given here, keep in mind three possible representations of observations on a real random variable:

(i) *Raw representation*. You write down the data in the order you observe them:

$$x_1, ..., x_n \tag{3.1}$$

(ii) *Ordered representation*. You put the observations on the real line and renumber them in an ascending order:

$$y_1 \le ... \le y_n \tag{3.2}$$

These are the same points as in (3.1), just the order is different, so the number of points is the same.

(iii) *Frequency representation*. This is a table similar to **Table 2.1**. That is, you see how many distinct values there are among the observed points and denote them $z_1 \le ... \le z_m$ (obviously, m does not exceed n). For each z_i you find its absolute frequency n_i and relative frequency r_i. The result will be

Table 3.1 Absolute and relative frequencies distribution

Values of the variable	Absolute frequencies	Relative frequencies
z_1	n_1	r_1
…	…	…
z_m	n_m	r_m
	Total = n	Total = 1

Our plan is to detail Figure 3.1 (below).

Unit 3.2   Measures of central tendency (sample mean, mean or average, median, mode; bimodal and trimodal distributions; outliers)

Mean. The *sample mean* (or simply *mean* or *average*) is the usual arithmetic mean

$$\bar{x} = \frac{x_1 + ... + x_n}{n} \tag{3.3}$$

Does it change if instead of the raw representation one uses the ordered representation?
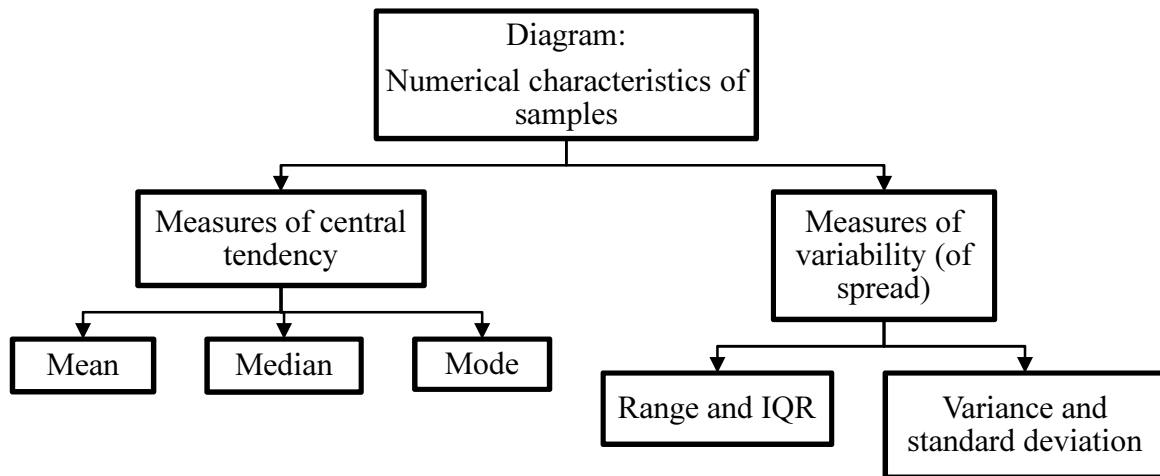
Figure 3.1 Numerical characteristics of samples

Median. **Intuitive definition**. The *median* is such a point that half of the observations lie to the left of it and another half to the right. It is useful to consider two samples (note that using the ordered representation is better).

Sample 1: $y_1 = 0, y_2 = 1, y_3 = 5$.   Sample 2: $y_1 = 0, y_2 = 1, y_3 = 3, y_4 = 5$.

In Sample 1, where n is odd, only $m=1$ satisfies the intuitive definition of the median. In Sample 2, where n is even, any point between 1 and 3 can serve as a median. To achieve uniqueness, it is customary to take 2, the middle of the interval $(1,3)$, as the median.

Before generalizing upon these examples, we have to think about a general form of writing even and odd numbers. You can verify that

(a) any number in the sequence 2, 4, 6, … can be written in the form $2k$ where k runs over naturals

and that

(b) any number in the sequence 1, 3, 5, … can be written in the form $2k+1$ where k runs over nonnegative integers.

Formal (and working) definition of the median.

Step 1. Arrange the observations in an ascending order.

Step 2. **(a)** If the number of observations is even, $n = 2k$, then the observations y_k and y_{k+1} will be the equivalents of the observations y_2 and y_3 in Sample 2 and

$$\text{median} = \frac{1}{2}(y_k + y_{k+1}).$$

(b) If the number of observations is odd, $n = 2k+1$, then the observation y_{k+1} will be the equivalent of the middle observation y_2 in Sample 1 and

$$\text{median} = y_{k+1}.$$

Mode. The *mode* is the most frequent observation. Obviously, in this case the frequency representation must be used. Note that it is possible for a frequency distribution to have more than one most frequent observation. In that case all of them will be modes. In the literature you can see names like *bimodal*, *trimodal distributions*.

Which of the measures of central tendency to use depends on the context and data.

Table 3.2 Exemplary comparison of measures of central tendency

Measure	Pros	Cons
Sample mean	Has the best theoretical properties (to be discussed in Unit 5.2 through Unit 5.5).	Can be applied only to numerical variables.
Median	Not sensitive to outliers. Good for income measurement.	Is not sensitive to where the bulk of the points lies.
Mode	Can be used for categorical variables. Good when only most frequent observation matters.	Insensitive to all but the most frequent observation.

Don't try to memorize this table. Think about the examples in NCT. Try to feel the definitions. Don't think about observations as something frozen. Try to move them around. For example, you can fix the median, and move the points to the left of it back and forth. As you do that, do the sample mean and mode change? Or you can think about dependence on *outliers*, which, by definition, are points which lie far away from the bulk of the observations. If you move them further away, which of the measures of central tendency may change and which do not?

Unit 3.3 Shape of the distribution (symmetry; positive and negative skewness; tails)

The definitions of symmetry and skewness must be given in terms of the histogram.

The distribution is called *symmetric* if the histogram is symmetric about the mean (that is, observations equidistant from the mean to the left and right must have equal frequencies), see Figure 10.21 in NCT. Empirical distributions rarely are exactly symmetric, and judgments about approximate symmetry are subjective.

We don't need the mathematical definition of *skewness*. The approximate definition of positive skewness given in NCT (that the mean should be greater than the median) is too often at variance with the precise definition. As a rule of thumb, it is better to talk about *tails*. We say that a distribution is *positively skewed* if the right tail is heavier than the left (think about the bars of the histogram as made of a heavy substance).

Unit 3.4 Measures of variability (range, quartiles, deciles, percentiles; interquartile range or IQR; five-number summary, sample variance, deviations from the mean, sample standard deviation)

Range. The smallest observation, denoted $\min x_i$, in terms of the ordered representation is, obviously, the leftmost point y_1. Similarly, the largest observation $\max x_i$ is the rightmost point y_n. Thus,

$$\min x_i = y_1, \ \max x_i = y_n.$$

(i) Sometimes people call by the *range* the difference $\text{Range} = y_n - y_1$.

(ii) It also makes sense to use this name for the segment $\text{Range} = [y_1, y_n]$. This is the smallest segment containing the sample.

Definition of quartiles, deciles, percentiles. The idea behind them is the same as with the median: certain intervals should contain certain fractions of the total number of observations. For example, in case of *quartiles* we want the picture in Figure 3.2 where each of the intervals $(Q_0, Q_1), (Q_1, Q_2), (Q_2, Q_3), (Q_3, Q_4)$ contains 25% of all observation.
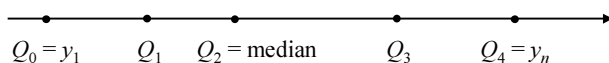
Figure 3.2 Five number summary

This leads to the following definition:

$$Q_1 = \text{observation numbered } 0.25(n+1)$$

$$Q_3 = \text{observation numbered } 0.75(n+1).$$

Actually, one has to take integer parts of the numbers $0.25(n+1)$, $0.75(n+1)$ to apply this definition. Similarly,

$$k\text{th } decile \text{ is the observation numbered } \frac{k}{10}(n+1)$$

$$k\text{th } percentile \text{ is the observation numbered } \frac{k}{100}(n+1).$$

In the literature you can see other definitions and none of them is perfect because all are approximate. Think about it: if you have just 20 observations, how can you divide them in 100 equal parts?

Definition of IQR. As explained in NCT, the range depends on outliers. When this is a problem, the spread of the middle 50% of the data may be preferable. This gives us the definition of the *interquartile range*, or *IQR*:

$$IQR = Q_3 - Q_1.$$

A five-number summary is the set of numbers in Figure 3.2:

$$\min x_i = Q_0 < Q_1 < Q_2 = median < Q_3 < Q_4 = \max x_i.$$

Variance. The *sample variance* is defined by

$$s_x^2 = \frac{1}{n-1}\left[(x_1 - \overline{x})^2 + ... + (x_n - \overline{x})^2\right]. \tag{3.4}$$

A few names will help you understand this construction. $x_1,...,x_n$ are the observed points. The differences $x_i - \overline{x}$ are called *deviations* (from the mean). $(x_i - \overline{x})^2$ are *squared deviations*. Their average would be

$$\frac{1}{n}\left[(x_1-\overline{x})^2+...+(x_n-\overline{x})^2\right].$$

In the definition of s_x^2 there is division by $n-1$ instead of n for reasons to be explained in Unit 7.2.

Exercise 3.1 If you can't tell directly from the definition how to calculate s_x^2, I insist that you do everything on the computer.

(i) If you are using the same Excel file as before, in cells D2-D21 you should have simulated observations on stock price. In cell G1 type `Mean` and in G2 enter the formula

```
=SUM($D$2:$D$21)/20
```

This is the average of the values $s_1,...,s_{20}$; the dollar signs will prevent the addresses from changing when you copy the formula to other cells.

(ii) Copy the formula to G2-G21. We want the same constant in G2-G21 to produce a horizontal line of the graph.

(iii) Select columns D, G and the line plot in the Chart Wizard. The time series plot is good to see the variability in the series. The sample mean is shown by a horizontal line. Some observations are above and others are below the mean, therefore some deviations from the mean are positive and others are negative.

(iv) In H1 type `Deviations`, in H2 type `=D2-G2` and copy this to H3-H21. In I1 type `Squared deviations`, in I2 enter `=H2^2` and copy this to I3-I21. You can notice that deviations have alternating signs and squared deviations are nonnegative.

(v) In H22 type `=SUM(H2:H21)/20` and in I22 enter the formula `=SUM(I2:I21)/19`. The value in H22, the average of deviations, will be close to zero showing that the mean deviation is not a good measure of variability. The value in I22 is s_x^2.

Standard deviation. The quantity $s_x=\sqrt{s_x^2}$ is called a *sample standard deviation*.

Lessons to be learned. Both sample variance and sample standard deviation are used to measure variability (or volatility, or spread). Both are based on deep mathematical ideas and for now you have to accept them as they are. Unlike the range or IQR, both depend on all observations in the sample.

Once I took a course in Financial Management in a Business Department. Several equations from Corporate Finance were used without explanations of the underlying theories. The whole course was about using formulas on the calculator. The Business students were happy and I was horrified. Read again the above procedure and make sure you see the forest behind the trees. You should be able to reproduce the definitions and, if necessary, use the calculator instead of Excel.

Exercise 3.2 (a) What happens to s_x^2 if all observations are increased by 5? Try to give two different explanations.

(b) Can the sample variance be negative?

(c) If the sample variance is zero, what can you say about the variable?

Unit 3.5 Measures of relationships between variables (sample covariance, sample correlation coefficient; positively

correlated, negatively correlated and uncorrelated variables; perfect correlation)

This is another complex topic which will be fully explained in Chapter 5.

A *sample covariance* is defined by

$$s_{xy} = \frac{1}{n-1}\left[(x_1 - \bar{x})(y_1 - \bar{y}) + ... + (x_n - \bar{x})(y_n - \bar{y})\right] \qquad (3.5)$$

where (x_i, y_i) are observations on a pair of variables. It is an auxiliary algebraic device that doesn't have much geometric or statistical meaning by itself. To remember the definition, compare it to s_x^2 and note that $s_{xx} = s_x^2$. After **Exercise 3.1** you should know how to calculate s_{xy}.

A *sample correlation coefficient* is defined by

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

where s_{xy} is the sample covariance and s_x, s_y are sample standard deviations.

These definitions are rooted in the Euclidean geometry. In particular, in Chapter 5 we shall see that

r_{xy} = cosine of the angle between two vectors associated with sample data.

For those who know properties of cosine this explains why the range of r_{xy} is the segment $[-1,1]$. We say that the variables x, y are *positively correlated* if $r_{xy} > 0$ and *negatively correlated* if $r_{xy} < 0$. In case $r_{xy} = 0$ the variables are called *uncorrelated*. In the extreme cases $r_{xy} = \pm 1$ we talk about *perfect correlation*. We use Excel to see what these names mean in terms of a scatterplot.

Exercise 3.3 (a) From **Exercise 3.1** it should be clear that calculating the sample correlation coefficient is a long story, even in Excel. We are going to take a shortcut using a built-in function CORREL. In **Exercise 2.6** we have plotted the scatterplot of the pair (Height, Weight). To find the correlation coefficient for the same pair, in cell J1 type Positive correlation and in cell J2 the formula
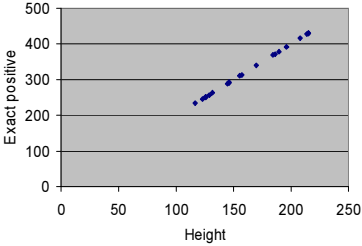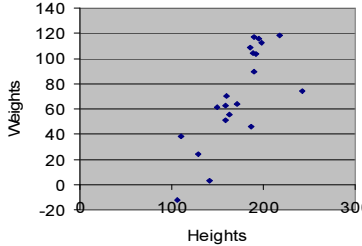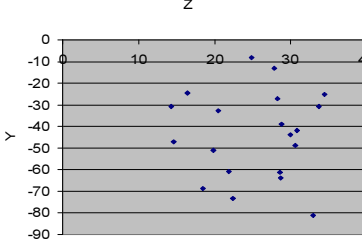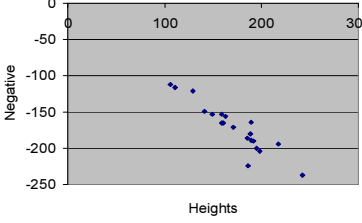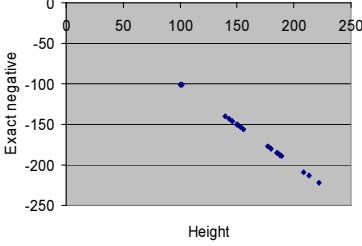
```
=CORREL(E2:E21,F2:F21)
```

(b) To generate a variable that is negatively correlated with Height, in K1 type Negative, in K2 put the formula =-E2+NORMINV(RAND(),0,20) and copy it to K3-K21. In L1 type Negative correlation and in L2 put the formula =CORREL(E2:E21,K2:K21). You know how to plot the scatterplot.

(c) For the table below you might want to arrange cases of perfect positive and negative correlations. Just use formulas like =2*E2 and =-E2. Simulation of uncorrelated variables can be done using **Exercise 12.3**(2).

Table 3.3 Summary of results on sample correlation coefficient

Values of r_{xy}	Statistical interpretation	Scatterplot

$r_{xy} = 1$	There is an exact linear relationship between y_i, x_i in the form $y_i = a + bx_i$ with a positive b.	
$0 < r_{xy} < 1$	There is an approximate linear relationship between y_i, x_i in the form $y_i \approx a + bx_i$ with a positive b.	
$r_{xy} = 0$	There is no definite pattern in the dependence.	
$-1 < r_{xy} < 0$	There is an approximate linear relationship between y_i, x_i in the form $y_i \approx a + bx_i$ with a negative b.	
$r_{xy} = -1$	There is an exact linear relationship between y_i, x_i in the form $y_i = a + bx_i$ with a negative b.	

Exercise 3.4 Form some pairs of variables: inflation, unemployment, aggregate income, investment, government expenditures, per capita income and birth rate. For each pair you form indicate the expected sign of the correlation coefficient.

Least-squares regression is one of important topics of this chapter. I am not discussing it because with the knowledge we have I cannot say about it more than the book, and what we have done should be enough for you to understand what the book says. End-of-chapter exercises 3.46-3.56 is the minimum you should be able to solve. The full theory is given in Chapter 12.

Unit 3.6   Questions for repetition

1. Give in one block all definitions related to the five-number summary

2. An expression for the sample variance equivalent to (3.4) is

$$s_X^2 = \frac{1}{n-1} \sum x_i^2 - \frac{1}{n(n-1)} \left(\sum x_i \right)^2.$$ Which one is easier to apply on a hand calculator?

3. Solve Exercise 3.45 from NCT.

Chapter 4   Probability
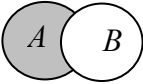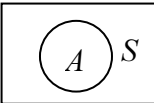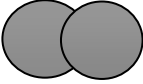
The way uncertainty is modeled in the theory of probabilities is complex: from sets to probabilities to random variables and their characteristics.

Unit 4.1   Set operations (set, element, union, intersection, difference, subset, complement, symmetric difference; disjoint or nonoverlapping sets, empty set)

Recall that the notion of a *set* is not defined. We just use equivalent names (collection, family, or array) and hope that, with practice, the right intuition will develop.

In the examples in the next table S denotes the set of all students of a particular university, A denotes the set of students taking an anthropology class and B is the set of students taking a biology class. We write $x \in A$ to mean that x is an *element of A* (*x belongs to A*). $x \notin A$ means that x is not an element of A.

Table 4.1 Set operations: definitions, visualization and logic

Definition	Visualization on Venn diagram	Logical interpretation and/or comments
The *union* $A \cup B$ is the set of elements which belong to A, to B or both.	Figure 4.2 in NCT	Logically, the union corresponds to nonexclusive "or"[7]. $A \cup B$ means all students who study anthropology or biology (or both).
The *intersection* $A \cap B$ is the set of elements common to A and B.	Figure 4.1(a) in NCT	Logically, the intersection corresponds "and". $A \cap B$ is the set of students who study both anthropology and biology.
The *difference* $A \setminus B$ is the set of elements of A which do not belong to B.		$A \setminus B$ is the set of students who study anthropology but not biology. Don't use "–" (minus) or "|" (pipe) instead of the backslash "\".
We say that A is a *subset* of S and write $A \subset S$ if all elements of A belong to S.		In applications S will be fixed, while its subsets will be changing.
A *complement* $\overline{A}$ is defined as the difference $\overline{A} = S \setminus A$ and includes elements of S outside A.	Figure 4.3 in NCT.	$\overline{A}$ is the set of students who do not study anthropology.
The set $A \triangle B = (A \setminus B) \cup (B \setminus A)$ is called a *symmetric difference* of A and B.	 (lightly shaded	The symmetric difference logically corresponds to "either or". $A \triangle B$ is the set of students who study A or B but not both.

---

[7] In English "or" is nonexclusive whereas "either or" is exclusive. In some other languages there is only one "or" and one has to specify whether it is exclusive or nonexclusive.

	area)	

For convenience, people introduce the *empty set* $\varnothing$ (the set without elements). A and B are called *disjoint* (or *nonoverlapping*) if $A \cap B = \varnothing$ (they have no common elements).

Unit 4.2 Set identities (distributive law, de Morgan laws, equality of sets)

To solve some exercises in NCT you need to use certain set relations. All of them can be established using Venn diagrams. There is no need to remember them but try to understand the logic which will help you pick the right relation for the exercise at hand.

Result 1. A set B can be obtained as a union of two disjoint sets, $B = (A \cap B) \cup (\overline{A} \cap B)$, where $A \cap B$ is the common part of A and B and $\overline{A} \cap B$ is the common part of $\overline{A}$ and B.

Try to express verbally the other results.

Result 2. $A \cup B = A \cup (\overline{A} \cap B)$.

Result 3. *Distributive law*: $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$.

Result 4. *De Morgan's laws*: $\overline{A \cup B} = \overline{A} \cap \overline{B}$, $\overline{A \cap B} = \overline{A} \cup \overline{B}$ (you pass to complements by interchanging unions and intersections).

These equations can be established using Venn diagrams but there is a more powerful method which works when using the Venn diagrams would be infeasible. We say that two sets, U and V, are equal, $U = V$, when $U \subset V$ and $V \subset U$ (that is each element of U belongs to V and vice versa). To prove the first de Morgan law, denote $U = \overline{A \cup B}$, $V = \overline{A} \cap \overline{B}$.

Suppose $x \in U$. By definition, $x \in S \setminus (A \cup B)$ which means two things: $x \in S$, $x \notin A \cup B$. Hence, $x \in \overline{A}$, $x \in \overline{B}$ and $x \in \overline{A} \cap \overline{B}$. Since the implication $x \in U \Rightarrow x \in V$ is true for any element of U, we have proved that $U \subset V$. The proof of $V \subset U$ is left as an exercise.

Unit 4.3 Correspondence between set terminology and probabilistic terminology (random experiment; impossible, collectively exhaustive and mutually exclusive events; sample space; basic outcomes or elementary events; occurrence of an event, disjoint coverings)

Definition of a random experiment: A *random experiment* is a process leading to two or more possible outcomes, with uncertainty as to which outcome will occur.

The example of rolling a die is sufficient to illustrate this intuitive definition.

> If you stay at the intuitive level, you will never really understand the true meaning of most mathematical constructions.

The next table helps you establish ways to think about outcomes. When more than one term exists to express the same thing, try to use the more geometric term.

Table 4.2 Set terminology and probabilistic terminology

	Set terminology	Probabilistic terminology
1	Set	*Outcome* or *event*, such as getting 1 on a die. Getting an even number is another example.

2	Empty set	*Impossible* event
3	Disjoint sets	*Mutually exclusive events*
4	S (universal set which contains all other sets of interest)	The widest possible event, called a *sample space S*. In case of the die, $S = \{1,2,3,4,5,6\}$
5	Sets $A_1,...,A_n \subset S$ form a *covering* if $S = A_1 \cup ... \cup A_n$.	Events $A_1,...,A_n \subset S$ such that $S = A_1 \cup ... \cup A_n$ are called *collectively exhaustive*.
6	Elements of S.	*Basic outcomes*, or *elementary events*, are the simplest possible events.

Comments. **(1)** In case of the die the basic outcomes are the sets $\{1\},\{2\},\{3\},\{4\},\{5\},\{6\}$. All other events can be obtained from these 6. For example, the event E that an even number turns up equals $E = \{2,4,6\}$. Of course, these numbers cannot turn up at the same time (this is expressed as $\{2\}\cap\{4\}\cap\{6\} = \varnothing$). We say that E *occurs* when one of the numbers 2, 4, 6 turns up.

(2) We distinguish a number 2 from a set $\{2\}$. Arithmetic operations apply to numbers, while for sets we use set operations.

In some problems we need *disjoint coverings*. For motivation, think about a jigsaw puzzle. Its pieces possess two properties: any two pieces are disjoint and together the pieces cover (comprise) the whole image. Departing from this geometric example, we say that events $A_1,...,A_n \subset S$ are *mutually exclusive* and *collectively exhaustive* if

(a) any pair A_i, A_j is disjoint and

(b) $S = A_1 \cup ... \cup A_n$.

Unit 4.4 Probability (inductive and deductive arguments; induction, probability, nonnegativity, additivity, completeness axiom, complement rule, assembly formula, impossible and sure events)

Before proceeding with the formal definitions, try to answer the following questions (in parentheses I give the answers I usually hear from my students).

Exercise 4.1 (a) Intuitively, what do you think is the probability of getting a head when tossing a coin? (Some say 0.5, others say 50%. There are also answers like: on average, 50 times out of 100).

(b) When rolling a die, what do you think is the probability of getting 1? (Most say 1/6).

(c) Recall **Table 2.1**. Suppose you obtain a large sample and calculate that the variable takes values listed in the first column with relative frequencies listed in the third column. For the purposes of predicting the values in future samples would you accept the relative frequencies as the probabilities of the respective values? (Most say "Yes").

The answers to parts (a) and (b) suggest that the coin and die are described by the following tables:

Table 4.3 Probability tables for C (Coin) and D (Die)

Values of C	Probabilities	Values of D	Probabilities
H	1/2	{1}	1/6
T	1/2	{2}	1/6

{3}	1/6
{4}	1/6
{5}	1/6
{6}	1/6

Denoting $P(A)$ the probability of event A, from **Table 4.3** we can surmise that

$$P(\{2,4,6\}) = P(\{2\}) + P(\{4\}) + P(\{6\}) = 3/6 = 1/2.$$

This makes sense because $\{1,2,3,4,5,6\} = \{1,3,5\} + \{2,4,6\}$. Generalizing upon **Table 2.1** and **Table 4.3**, in case of a variable with n values the following should be true:

Table 4.4 Probability table in case of a variable with n values

Values of the variable	Probabilities
x_1	$p_1 = P(X = x_1)$
…	…
x_n	$p_n = P(X = x_n)$

Here the sample space is $S = \{x_1, ..., x_n\}$. Since the numbers p_i mean percentages, they should satisfy the conditions

   (i) $0 < p_i < 1$ for all i and

   (ii) $p_1 + ... + p_n = 1$ (*completeness axiom*).

The completeness axiom means that we have listed in the table all possible basic outcomes. If that axiom were not satisfied, we would have a smaller sample space.

Until this point I have been using what is called an *inductive argument*: by analyzing a simple situation and using an element of guessing, try to come up with definition that could be a base of a theory. When a definition is difficult to understand, try to pinpoint the underlying inductive argument. In the times of Karl Gauss it was common to show inductive arguments. These days they are mostly omitted, partly to save on paper and time. By skipping inductive arguments, you are not saving your time. You are complicating your task.

Before turning to a *deductive argument*, in which everything is deduced logically from a few basic definitions and postulates, a brush up on functions is necessary. Whenever you talk about a function $f(x)$, you have to realize what type of objects you can substitute as arguments and what type of objects you can get as values. Right now we need to discuss dependence of area on a surface and of volume on a body. In both cases the arguments are sets and the values are numbers. Most importantly, the functions are additive. For instance, if we denote x_i pieces of a jigsaw puzzle, then

$$area(x_1 \cup ... \cup x_n) = area(x_1) + ... + area(x_n).$$

Similarly, if r_i denotes a room in a building, then

$$volume(r_1 \cup ... \cup r_n) = volume(r_1) + ... + volume(r_n).$$

Note the importance, for these equations to hold, of the facts that the pieces $x_1, ..., x_n$ do not overlap and the rooms $r_1, ..., r_n$ are disjoint. Put it simply, if $A \cap B = \varnothing$, as in the diagram
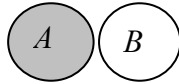
then $area(A \cup B) = area(A) + area(B)$. If, on the other hand, $A \cap B \neq \varnothing$, as in the more general case
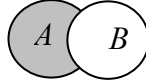
then $area(A \cup B) = area(A) + area(B) - area(A \cap B)$ because in the sum $area(A) + area(B)$ the area of $A \cap B$ is counted twice.

Definition. Let S be a sample space. By *probability* on S we mean a numerical function $P(A)$ of events $A \subset S$ such that

(a) if A, B are mutually exclusive, then $P(A \cup B) = P(A) + P(B)$ (*additivity*),

(b) for any A, $P(A)$ is *nonnegative*, and

(c) $P(S) = 1$ (*completeness axiom*).

Any definition as complex as this needs some chewing and digesting. This includes *going back*, to the motivating examples, *going forward*, to the consequences, and *going sideways*, to look at variations or cases when it is not satisfied.

Going back. To put the example from **Table 4.4** into the general framework, for any event $A \subset S$ define

$$P(A) = \sum_{x_i \in A} P(\{x_i\}) \quad \text{(read like this: sum of probabilities of those}$$

$$x_i \text{ which belong to } A).$$

For example, $P(\{x_1, x_2, x_3\}) = P(\{x_1\}) + P(\{x_2\}) + P(\{x_3\})$. Later on we are going to use more summation signs, and you will get used to them. Always write out the sums completely if you don't understand summation signs. As an exercise, try to show that the $P(A)$ defined here is additive, nonnegative and satisfies the completeness axiom.

Going forward. **(1)** One way to understand a property is to generalize it. In the case under consideration, additivity is generalized to the case of n events. If $A_1, ..., A_n$ are disjoint, then from item (a) in the definition of probability we have

$$P(A_1 \cup ... \cup A_n) = P((A_1 \cup ... \cup A_{n-1}) \cup A_n) = P(A_1 \cup ... \cup A_{n-1}) + P(A_n)$$
$$= P(A_1 \cup ... \cup A_{n-2}) + P(A_{n-1}) + P(A_n) = P(A_1) + ... + P(A_n).$$

This type of derivation of a general statement from its special case is called *induction*.

(2) The complement rule. Representing S as $S = A \cup \overline{A}$, from items (a) and (c) we deduce

$$P(A) + P(\overline{A}) = P(S) = 1 \tag{4.1}$$

which implies the *complement rule*:

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

(3) From (b) we know that $P(A) \geq 0$. In addition, since in (4.1) $P(\overline{A}) \geq 0$, we have $P(A) \leq 1$.

(4) *Assembly formula*. Fix some event B. If $A_1, ..., A_n$ are disjoint, then pieces $B \cap A_1, ..., B \cap A_n$ are also disjoint. Further, if $A_1, ..., A_n$ are collectively exhaustive, then those pieces comprise B. Therefore by additivity

$$P(B) = P(B \cap A_1) + ... + P(B \cap A_n). \tag{4.2}$$

Going sideways. While area and volume are additive and nonnegative, they generally do not satisfy condition (c).

When a definition is long, after chewing and digesting it you should formulate its short-and-easy-to-remember version.

Short definition. A *probability* is a nonnegative additive function defined on some sample space S and such that $P(S) = 1$.

An *impossible event* can be defined by $P(A) = 0$. When $P(A) = 1$, we say that A is a *sure event*. In the discrete case we are considering the only impossible event is $A = \varnothing$ and the only sure event is the sample space.

In many practical applications it is not necessary or possible to check that all requirements of probabilistic definitions are satisfied.

Exercise 4.2 (a) What is the probability that tomorrow the sun will rise?

(b) What is the probability of you being at home and at the university at the same time?

(c) Probability of which event is higher: that tomorrow's temperature at noon will not exceed 19°C or that it will not exceed 25°C?

Unit 4.5   Ways to find probabilities for a given experiment (equally likely events, classical probability, addition rule)

4.5.1 Equally likely outcomes (theoretical approach)

(1) In case of a fair coin, the head and tail are *equally likely*, so $P(H) = P(T)$. Since these events are mutually exclusive and collectively exhaustive, by additivity and completeness axiom $P(H) + P(H) = 1$. It follows that $P(H) = P(T) = 1/2$. We have rigorously proved the first Table 4.3.

(2) In case of a fair die the events $\{1\}, ..., \{6\}$ are equally likely, so they all have the same probability p. By additivity and completeness axiom $6p = 1$, so $p = 1/6$. We have derived the second Table 4.3.

(3) More generally, suppose that there are N basic outcomes and suppose the basic outcomes are equally likely. Then, as above, the probability of one basic outcome is $1/N$. Now consider some condition that separates an event $A \subset S$ (in case of a die the condition could be that the outcome is even). Denote N_A the number of basic outcomes satisfying the condition. Then by additivity

$$P(A) = \underbrace{\frac{1}{N} + ... + \frac{1}{N}}_{N_A \text{ times}} = \frac{N_A}{N}.$$

This formula is called *classical probability* in the book. Many students remember the formula but don't think about the conditions under which it is true. Reread the above derivation and state the result in the form: if … then…

(4) Frequentist approach. After observing relative frequencies in an experiment, one can take those frequencies as probabilities. The larger the sample size, the better will be the approximation to the truth. This is a general statistical fact. Statistics is mostly about large sample properties.

Exercise 4.3 We are going to simulate a large number of tossings of a fair coin. To work with numbers instead of heads and tails, we introduce a numerical variable

$$C = \begin{cases} 1 \text{ if head} \\ 0 \text{ if tail} \end{cases}$$

In a large number of trials n we calculate the proportion of heads. As n increases, p_n should approach ½.

Step 1. In cell A1 put the formula

$$=IF(RAND()>1/2,0,1)$$

Copy it to cells A2-A100. These will be observations on the coin.

Step 2. In B1 type

$$=SUM(\$A\$1:A1)/ROW(A1)$$

and copy this formula to B2-B100. The result in, say, B100 will be

$$=SUM(\$A\$1:A100)/ROW(A100)$$

This is the sum of the contents of the cells A1 through A100 divided by the number of these cells. Thus, in cells B1-B100 we have averages of the first n observations, $n = 1,...,100$.

Step 3. Select column B and see on the line plot the behavior of averages. They approach ½. To see the picture better, on the plot right-click the vertical axis and select "Format Axis". On the Scale tab check Minimum and enter 0 for the value, uncheck Maximum and enter 1 for the value and check Major Unit and enter 0.1 for the value. Click OK. By pressing F9 you will change realizations of the means for sample sizes $n \leq 100$. Sometimes the speed of convergence of averages to ½ is not very good. This is a general probabilistic fact (see Unit 8.2 for the theory).

(5) Expert opinions. When the theoretical and frequentist approaches are not applicable, one can use subjective probabilities assigned by experts. For example, performance of economic projects depends on overall macroeconomic conditions. In project evaluation the resulting statements sound like this: with probability 0.6 the economy will be on the rise, and then the profitability of the project will be this much and so on.

Exercise 4.4 Generalize the *addition rule* $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ to the case of three events $P(A \cup B \cup C) = ...$

Unit 4.6   Combinatorics (factorial, orderings, combinations)

Definition. For a natural number n denote $n! = n \cdot (n-1)...2 \cdot 1$. This number is called n *factorial*.

What do you think grows faster (as n increases): e^n, $n!$ or n^n?

Combinatorics is a wide area but we need only two results, one about orderings and the other about combinations. In order not to confuse them, remember just two motivating examples.

Example 4.1 There are three contestants in a competition, denoted A_1, A_2, A_3, and three prizes of different values (the possibility of ties is excluded). How many prize distributions are possible?

Solution. Let ">" stand for "better". The ordering $A_1 > A_2 > A_3$ is not the same as, say, $A_2 > A_1 > A_3$ because the prizes are different. The number of prize distributions equals the number of orderings and can be found as follows. Any of the three contestants can get the first prize, which gives us 3 possibilities. After the first prize is awarded, any of the remaining two contestants can get the second prize, which gives us 2 possibilities. After the second prize is awarded, there is only one possibility for the third prize. Any of these possibilities (3, 2 and 1) can be combined with any other, so the total number of possibilities is $3 \cdot 2 \cdot 1 = 3!$ To see why the possibilities must be multiplied, you can use a tree (see NCT).

Exercise 4.5 To internalize the above argument, consider the case of n contestants and n prizes. The number of *orderings* is the number of ways n elements can be ordered and it equals $n!$.

Example 4.2 There are three new professors on campus and the IT department has 5 new identical computers, 3 of which will be distributed among the new professors. In how many different ways this is possible to do?

Solution. Denote A_1, A_2, A_3 the professors and $C_1, C_2, ..., C_5$ the computers. Since the computers are identical, the computer assignment

$$C_1 \to A_1$$
$$C_2 \to A_2$$
$$C_3 \to A_3$$

is as good as, say,

$$C_1 \to A_1$$
$$C_3 \to A_2 \ .$$
$$C_5 \to A_3$$

To make use of **Example 4.1** let us introduce two fictitious persons F_1, F_2. As we know, A_1, A_2, A_3, F_1, F_2 can be assigned computers in 5! different ways, if the order matters. Now we take into account that real persons can be ordered in 3! ways. Fictitious persons can be ordered in 2! ways. Five people can be ordered in 3!2! ways. Thus, the number of computer assignments to real persons when the order doesn't matter is $\dfrac{5!}{3!2!}$

Definition. The number of different choices of x objects out of n identical objects is called the *number of combinations of x elements out of n* and denoted C_x^n (here $0 \le x \le n$).

Exercise 4.6 Generalize the above argument to show that $C_x^n = \dfrac{n!}{x!(n-x)!}$

Comments. **(1)** When the order matters, use orderings and when it does not, apply combinations.

(2) In case of combinations I find it more practical to read C_x^n as "the number of choices of x elements out of n".

(3) For the choices formula to work in the extreme cases $x = 0, x = n$ we put formally $0! = 1$.

Exercise 4.7 Write out sequences $\{C_0^1, C_1^1\}$, $\{C_0^2, C_1^2, C_2^2\}$, $\{C_0^3, C_1^3, C_2^3, C_3^3\}$. Note that they first grow and then decline, the first and last elements being 1.

Unit 4.7 All about joint events (joint events, joint and marginal probabilities, cross-table, contingency table)

These notions are interrelated but in NCT they are given in different places and without discussing the relationships between them. The discussion will be long, and if you are stuck with something, read everything from the start.

Example 4.3 In Kazakhstan the profits of wheat producers critically depend on weather conditions during summer (a lot of rain results in an abundant crop; there is always a lot of sun) and the weather during the harvesting time (just three rainy days in September are enough for the quality of wheat to plunge). Let us denote by S the event that the weather is good in summer and by H the event that it is good during the harvesting time. The farmers are happy when $P(S \cap H)$ is high.

Definition. The event $A \cap B$ is called a *joint event* and its probability $P(A \cap B)$ is called a *joint probability*.

Comments. **(1)** Following the recommendation made earlier, remember that A and B may vary and $P(A \cap B)$ is a function of two sets.

(2) In the wheat example we have four possible states of nature (the basic events) $H \cap S$, $H \cap \bar{S}$, $\bar{H} \cap S$, $\bar{H} \cap \bar{S}$ and, correspondingly, four probabilities $P(H \cap S)$, $P(H \cap \bar{S})$, $P(\bar{H} \cap S)$, $P(\bar{H} \cap \bar{S})$. By the completeness axiom, these probabilities sum to 1. As usual, this information can be displayed in a table.

Table 4.5 1-D table of joint events and probabilities

Events	Probabilities
$H \cap S$	$P(H \cap S)$
$H \cap \bar{S}$	$P(H \cap \bar{S})$
$\bar{H} \cap S$	$P(\bar{H} \cap S)$
$\bar{H} \cap \bar{S}$	$P(\bar{H} \cap \bar{S})$
	Total = 1

Note that by the complement rule $P(S) + P(\bar{S}) = 1$, $P(H) + P(\bar{H}) = 1$. Therefore we have two more tables

Table 4.6 Separate probability tables

Events	Probabilities		Events	Probabilities
S	$P(S)$		H	$P(H)$
$\bar{S}$	$P(\bar{S})$		$\bar{H}$	$P(\bar{H})$
	1			1

It is convenient to call the numbers $P(S)$ and $P(\overline{S})$ *own probabilities* of the pair $\{S, \overline{S}\}$ and the numbers $P(H)$ and $P(\overline{H})$ *own probabilities* of the pair $\{H, \overline{H}\}$.

(3) It is better to replace a 1-D table **Table 4.5** by a 2-D table, in order to see better its link to **Table 4.6**.

Table 4.7 Joint probability table

	H	$\overline{H}$	Right margin
S	$P(S \cap H)$	$P(S \cap \overline{H})$	$P(S)$
$\overline{S}$	$P(\overline{S} \cap H)$	$P(\overline{S} \cap \overline{H})$	$P(\overline{S})$
Lower margin	$P(H)$	$P(\overline{H})$	1

In a 2-D table it is easier to form pairs of individual events S, $\overline{S}$, H and $\overline{H}$. The probabilities from **Table 4.5** have migrated to the central part of **Table 4.7** (circumscribed with a double line). By the assembly formula

$$P(S \cap H) + P(S \cap \overline{H}) = P(S), \; P(\overline{S} \cap H) + P(\overline{S} \cap \overline{H}) = P(\overline{S}).$$

Thus, row sums of the probabilities in the central part of **Table 4.7** are equal to own probabilities of the pair $\{S, \overline{S}\}$ (displayed in the right margin). Similarly, column sums of the probabilities in the central part of **Table 4.7** are equal to own probabilities of the pair $\{H, \overline{H}\}$ (displayed in the lower margin). The customary name for own probabilities, *marginal probabilities*, reflects their placement in the table but not their role.

(4) The unity in the lower right corner of **Table 4.7** is

      (a) the sum of joint probabilities in the central part,

      (b) the sum of own probabilities in the right margin and

      (c) the sum of own probabilities in the lower margin.

(5) A *cross-table* (also called a *contingency table*) is similar to the joint probabilities table, except that instead of probabilities absolute frequencies are used.

Exercise 4.8 (a) List and prove 7 identities which hold for a joint probabilities table.

(b) Draw a joint probabilities table for the case when in the first column there are events A_1, …, A_n and in the first row there are events B_1, …, B_m (n is not necessarily equal to m).

Summation signs. Because of importance of the last exercise, we'll do it together. This is the place to learn the techniques of working with summation signs. The sum

$$a_1 + ... + a_n \tag{4.3}$$

(in which the three points mean "and so on") is written as

$$\sum_{i=1}^{n} a_i \tag{4.4}$$

(4.3) can be referred to as an *extended expression* and (4.4) is a *short expression*. In the short expression we indicate the index that changes, i, and that it runs from the *lower limit of summation*, 1, to the *upper* one, n. If you have problems working with summation signs, replace them with extended expressions, do the algebra, and then convert everything back. For the beginning remember two things:

(i) The sum (4.4) does not depend on i, as one can see from the extended expression. In particular, the summation index can be replaced by anything else, for example, $\sum_{j=1}^{n} a_j$ is the same thing as (4.4).

(ii) When more than one different indices are involved, it is better to use different notations for them.

The answer to **Exercise 4.8**(b) looks as follows:

Table 4.8 Joint probability table (general case)

	B_1	...	B_m	Right margin
A_1	$P(A_1 \cap B_1)$	...	$P(A_1 \cap B_m)$	$P(A_1)$
...	...	...	...	...
A_n	$P(A_n \cap B_1)$	...	$P(A_n \cap B_m)$	$P(A_n)$
Lower margin	$P(B_1)$	...	$P(B_m)$	1

Each of the families $\{A_1, \ldots, A_n\}$ and $\{B_1, \ldots, B_m\}$ is mutually exclusive and collectively exhaustive.

Identities:

(i) Sums across rows: $\sum_{j=1}^{m} P(A_1 \cap B_j) = P(A_1), \ldots, \sum_{j=1}^{m} P(A_n \cap B_j) = P(A_n)$.

(ii) Sums across columns: $\sum_{i=1}^{n} P(A_i \cap B_j) = P(B_j), \ j = 1,\ldots,m.$

(iii) Sums of own probabilities: $\sum_{j=1}^{m} P(B_j) = 1, \sum_{i=1}^{n} P(A_i) = 1.$

(iv) Sum of joint probabilities: $\sum_{j=1}^{m} \sum_{i=1}^{n} P(A_i \cap B_j) = 1.$

Unit 4.8 Conditional probabilities (multiplication rule, independence of events, prior and posterior probability)

Example 4.4 In the wheat example (**Example 4.3**), suppose the summer is over, the weather has been good and the harvesting time is approaching. We are left with one line of **Table 4.7**:

Table 4.9 Leftover joint probabilities

	H	$\bar{H}$	Right margin
S	$P(S \cap H)$	$P(S \cap \bar{H})$	$P(S)$

Here the sum

$$P(S \cap H) + P(S \cap \bar{H}) = P(S) \qquad (4.5)$$

is not 1 and the completeness axiom is not satisfied. But we know that there are no missing events. To satisfy the completeness axiom, we can divide both sides of (4.5) by $P(S)$,

$$\frac{P(S \cap H)}{P(S)} + \frac{P(S \cap \bar{H})}{P(S)} = 1$$

and treat $P(S \cap H)/P(S)$ and $P(S \cap \bar{H})/P(S)$ as probabilities of events H and $\bar{H}$, provided that S has occurred.

Definition. Let $P(S) > 0$. The probability $P(H \mid S) = \dfrac{P(H \cap S)}{P(S)}$ is called *probability of* H *conditional on* S.

Comments. (1) If we are interested in probabilities of H and $\bar{H}$, why not use simply $P(H)$ and $P(\bar{H})$? The reason is that occurrence of S provides us with additional information. There may be some dependence between events S and H, in which case conditional probabilities $P(H \mid S)$ and $P(\bar{H} \mid S)$ will be a better reflection of reality than own probabilities of the pair $\{ H , \bar{H} \}$. This process of updating our beliefs as new information arrives is one of the fundamental ideas of the theory of probabilities.

(2) Often random variables are introduced axiomatically. That is, take any numbers p_i satisfying $0 < p_i < 1$ and $p_1 + ... + p_n = 1$ and you can be sure that there is a random variable whose values have probabilities exactly equal to those p_i's. Conditional probabilities are defined axiomatically and the subsequent theory justifies their introduction.

Multiplication rule. The definition of $P(H \mid S)$ immediately implies

$$P(H \cap S) = P(H \mid S)P(S).$$

Try to see the probabilistic meaning of this equation. Denote M the event that you obtain a passing grade in Math and by S the event that you obtain a passing grade in Stats. Suppose that Math is a prerequisite for Stats. Then the equation $P(S \cap M) = P(S \mid M)P(M)$ basically tells us that to obtain a passing grade in both Math and Stats you have to pass Math first and, with that prerequisite satisfied, to obtain a passing grade in Stats. The fact that you have not failed Math tells something about your abilities and increases the chance of passing Stats.

4.8.1 Independence of events

Intuitive definition. Events A and B are called *independent* if occurrence of one of them does not influence in any way the chances of occurring of the other. For example, what happens to the coin has nothing to do with what happens to the die.

Formal definition. Events A and B are called *independent* if $P(A \cap B) = P(A)P(B)$. When $P(B) > 0$, this can be equivalently written as $P(A \mid B) = P(A)$.

This definition is applied in two ways.

(1) From independence to equation. If you know that A and B are independent, you can multiply $P(A)$ and $P(B)$ to get $P(A \cap B)$. For example,

$$P(Coin = 1, Die = 1) = P(Coin = 1)P(Die = 1) = 1/12.$$

(2) From equation to independence (or dependence). If you find that $P(A \cap B) \neq P(A)P(B)$, you conclude that A and B are dependent.

Exercise 4.9 Prove that if A and B are independent, then their complements are also independent.

Solution. Always write down what is given and what you need to get. We know that $P(A \cap B) = P(A)P(B)$ and we need to prove that $P(\overline{A} \cap \overline{B}) = P(\overline{A})P(\overline{B})$. When rearranging expressions indicate what rule you are applying, until everything becomes trivial.

$$
\begin{aligned}
P(\overline{A})P(\overline{B}) = & \qquad \text{(complement rule)} \\
= & [1 - P(A)][1 - P(B)] \qquad \text{(multiplying out)} \\
= & 1 - P(A) - P(B) + P(A)P(B) \quad \text{(by independence)} \\
= & 1 - P(A) - P(B) + P(A \cap B) \quad \text{(additivity rule)} \\
= & 1 - P(A \cup B) \qquad \text{(complement rule)} \\
= & P(\overline{A \cup B}) \qquad \text{(de Morgan)} \\
= & P(\overline{A} \cap \overline{B})
\end{aligned}
$$

A tangible example of independent events

Take the square $S = [0,1] \times [0,1]$ as the sample space, see Figure 4.1. For any set $A \subset S$ define its probability to be its area: $P(A) = area(A)$. Since area is additive, nonnegative and $P(S) = 1$, this will be really a probability.
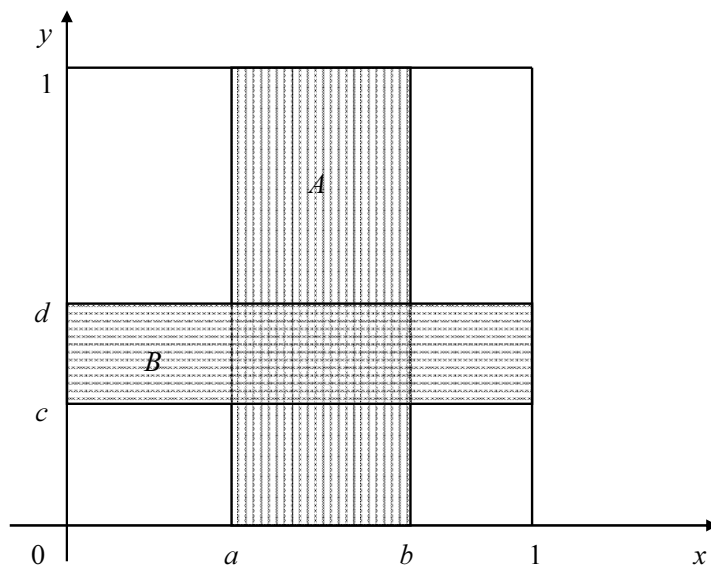
Figure 4.1 Example of independent events

For any $[a,b] \subset [0,1]$ define A as the vertical strip $A = \{(x,y) : a \leq x \leq b, 0 \leq y \leq 1\}$. For any $[c,d] \subset [0,1]$ define B as the horizontal strip $B = \{(x,y) : 0 \leq x \leq 1, c \leq y \leq d\}$. Then

$$
\begin{aligned}
& P(A) = b - a, \ P(B) = d - c, \\
& A \cap B = \{(x,y) : a \leq x \leq b, c \leq y \leq d\}, \\
& P(A \cap B) = (b-a)(d-c) = P(A)P(B)
\end{aligned}
$$

Thus, A and B are independent. Now try to move the strip A left and right and the strip B up and down. Independence takes a physical meaning: movements in mutually orthogonal directions do not affect each other.

Theorem 4.1 (*Bayes theorem*) This theorem is one of realizations of the idea about updating one's beliefs on the basis of new information. The easiest way to remember it is to express $P(B \mid A)$ through $P(A \mid B)$:

$$P(B \mid A) = \frac{P(B \cap A)}{P(A)} = \frac{P(B \cap A)}{P(B)} \frac{P(B)}{P(A)} = P(A \mid B) \frac{P(B)}{P(A)}.$$

Interpretation in terms of subjective probabilities. Before occurrence of the event A, an expert forms an opinion about likelihood of the event B. That opinion, embodied in $P(B)$, is called a *prior probability*. After occurrence of A, the expert's belief is updated to obtain $P(B \mid A)$, called a *posterior probability*. By the Bayes theorem, updating is accomplished through multiplication of the prior probability by the factor $\frac{P(A \mid B)}{P(A)}$.

A simple way to complicate the matters. Suppose $B_1, ..., B_K$ form a disjoint covering of the sample space. By the Bayes theorem for each B_i

$$P(B_i \mid A) = \frac{P(A \mid B_i) P(B_i)}{P(A)}.$$

Here $P(A)$ can be replaced by

$$
\begin{aligned}
P(A) &= & \text{(assembly formula)} \\
&= P(A \cap B_1) + ... + P(A \cap B_K) & \text{(multiplication rule)} \\
&= P(A \mid B_1) P(B_1) + ... + P(A \mid B_K) P(B_K)
\end{aligned}
$$

The result is

$$P(B_i \mid A) = \frac{P(A \mid B_i) P(B_i)}{P(A \mid B_1) P(B_1) + ... + P(A \mid B_K) P(B_K)}. \tag{4.6}$$

To me this is a good proof that knowing the derivation is better than remembering the result.

Unit 4.9   Problem solving strategy

Make sure that you can solve at least two typical exercises from each section. Have a look at Exercises (in NCT) 4.5, 4.6, 4.13, 4.14, 4.27, 4.28, 4.33, 4.35, 4.66, 4.76, 4.85, and 4.86. In the end of each chapter, starting from this one, there is a unit called Questions for Repetition, where, in addition to theoretical questions, I indicate a minimum of exercises from NCT.

To write down the conditions, you need a good notation. Avoid abstract notation and use something that reminds you of what you are dealing with, like A for Anthropology class.

When looking for a solution of a problem, write down all the relevant theoretical facts. Don't limit yourself to conclusions; repeat everything with proofs and derivations. Omit only what is easy; after a while everything will become easy. If you still don't see the solution, go back by one or two steps and solve simpler similar exercises.

Exercises in the end of Sections 4.4 and 4.5 are easier than they look. This is where writing down all the relevant equations is particularly useful. Remember what you need to find. In general, to find n unknowns it is sufficient to have n equations involving them. One of my students was able to solve those exercises without any knowledge of conditional and joint probabilities, just using common sense and the percentage interpretation of probabilities.

When you clearly see the idea, don't calculate the numbers. Learn to save your time by concentrating on ideas.

If section exercises are too simple, don't waste your time and go directly to end-of-chapter exercises.

Everybody finds difficult exercises that require orderings and combinations (there are many of them in the end of Section 4.3 and Chapter 4). Solve more of them, until you feel comfortable. However, it's not a good idea to solve all of them.

> Don't go wide, go deep!

Unit 4.10   Questions for repetition

1. What do we mean by set operations?

2. Prove de Morgan laws.

3. Prove and illustrate on a Venn diagram equation (4.2).

4. Prove $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

5. Prove the formulas for orderings and combinations

6. Using a simple example, in one block give all definitions and identities related to joint and marginal probabilities.

7. Solve **Exercise 4.9**.

8. Derive the complex form of the Bayes theorem (4.6).

9. The minimum required: Exercises 4.5, 4.13, 4.27, 4.37, 4.47, 4.67, 4.88, 4.91, 4.93, 4.97, 4.107 from NCT.

Chapter 5   Discrete Random Variables and Probability Distributions

At some point you may feel that the amount of new information is overwhelming and you forget many facts. Then it's the time to arrange a global repetition. Repeat everything with proofs. Look for those minor or significant details you missed earlier. Make a list of facts in a logical order. This chapter will be more challenging than all the previous taken together, and you will need to be in a perfect shape to conquer it.

Unit 5.1   Random variable (random variable, discrete random variable)

Intuitive definition. A *random variable* is a variable whose values cannot be predicted.

For example, we can associate random variables with the coin and die. On the other hand, to know the value of $\sin x$ it suffices to plug the argument x in. $\sin x$ or the number of days in a year are not random variables.

Short formal definition. A *random variable* is a pair: values plus probabilities.

Working definition of a discrete random variable. A *discrete random variable* is defined by Table 5.1:

Table 5.1 Discrete random variable with n values

Values of X	Probabilities
x_1	p_1
…	…
x_n	p_n

Comments. **(1)** Most of the time, when people talk about random variables, they assume real-valued variables. Thus, in the first column we have real numbers. This is important to be able to use arithmetic operations.

(2) Capital X is used for the variable and lower-case x's for its values. The complete form of writing p_i is $P(X = x_i)$.

(3) Without the second column we would have a deterministic variable.

(4) In the table X takes a finite number of values. This is a simplification. In Unit 5.14 we'll see an example of a discrete random variable that takes an infinite number of values.

Unit 5.2   General properties of means (expected value, mean, average, mathematical expectation, uniformly distributed discrete variable, grouped data formula, weighted mean formula)

Example 5.1 With probability 0.001 a strong earthquake destroys your house this year and then your loss is $200,000. Otherwise, your loss is just normal wear and tear, estimated at $4,000 per year. What is your expected loss?

Sometimes I ask questions before giving formal definitions, to prompt you to guess. Guessing and inductive reasoning are important parts of our thinking process. In economics there is a French term "tâtonnement" (groping one's way in the dark). There are cases when you

clearly see the idea and the solution is straightforward. In more complex situations there may be several competing ideas, and some of them may not be clear, looking like unfamiliar creatures in the dark. Looking for them and nurturing them until they shape up so that you can test them is a great skill. This is where you need to collect all the relevant information, mull it over again and again, until it becomes your second nature. This process is more time-consuming than getting a ready answer from your professor or classmate but the rewards are also greater.

The expected loss is $(-4,000)\times0.999 + (-200,000)\times0.001 = -4,196$. The general idea is that it is a weighted sum of the values of the variable of interest (the loss function). We attach a higher importance to the value that has a higher probability.

> In research ideas go before definitions.

Expected value of a discrete random variable. Using the working definition of a random variable, we define the *expected value* of X (also called a *mean*, or an *average*, or *mathematical expectation*) to be

$$EX = x_1 p_1 + ... + x_n p_n. \tag{5.1}$$

Comments. **(1)** Expected value is a function whose argument is a complex object (it is described by Table 5.1) and the value is simple: EX is just a number. And it is not a product of E and X !

(2) Relation to the sample mean. A discrete random variable is said to be *uniformly distributed* if $p_1 = ... = p_n = 1/n$. In this case the expected value is just the sample mean:

$$EX = x_1 \frac{1}{n} + ... + x_n \frac{1}{n} = \frac{x_1 + ... + x_n}{n} = \overline{X}.$$

(3) Grouped data formula. Recall the procedure for finding relative frequencies. Denote $y_1,..., y_n$ the values in the sample. Equal values are joined in groups. Let $x_1,...,x_m$ denote the distinct values and $n_1,...,n_m$ their absolute frequencies. Their total is, clearly, n. The sample mean is

$$\overline{Y} = \frac{y_1 + ... + y_n}{n} \qquad \text{(sorting out } y\text{'s into groups with equal values)}$$

$$= \frac{\overbrace{x_1 + ... + x_1}^{n_1 \text{ times}} + ... + \overbrace{x_m + ... + x_m}^{n_m \text{ times}}}{n} \tag{5.2}$$

$$= \frac{n_1 x_1 + ... + n_m x_m}{n} \qquad \text{(dividing through by } n \text{ and using relative frequencies)}$$

$$= r_1 x_1 + ... + r_m x_m = EX.$$

In this context, the sample mean $\overline{Y}$, the *grouped data formula* (in the third line) and the expected value are the same thing. The *weighted mean formula* from NCT (see their equation (3.12)) is the same thing as the grouped data formula with $n_i = w_i$ and $n = \sum w_i$. Note also that when data are grouped by intervals, as in Table 3.5 of NCT, in the grouped data formula the midpoints m_i of the intervals are taken as the values of the variables.

(4) When the sample size becomes very large, the relative frequencies approach the true probabilities. Therefore (5.2) gives rise to the following intuitive interpretation of the expected value: EX is what $\overline{X}$ approaches when n is very large.

Unit 5.3   Linear combinations of random variables (linear combination, vector, parallelogram rule)

Definition. Suppose that X, Y are two discrete random variables with the same probability distribution $p_1, ..., p_n$. Let a, b be real numbers. The random variable $aX + bY$ is called a *linear combination* of X, Y with coefficients a, b. Its special cases are aX (X *scaled* by a) and $X + Y$ (a *sum* of X and Y). The detailed definition is given by the next table.

Table 5.2 Table of linear operations with random variables

Values of X	Values of Y	Probabilities	aX	$X + Y$	$aX + bY$
x_1	y_1	p_1	ax_1	$x_1 + y_1$	$ax_1 + by_1$
…	…	…	…	…	…
x_n	y_n	p_n	ax_n	$x_n + y_n$	$ax_n + by_n$

Comments. **(1)** There is only one column of probabilities because X and Y are assumed to have the same probability distribution. The situation when the probability distributions are different is handled using joint probabilities (see Unit 5.5).

(2) If you are asked what are vectors, you say the following. By a *vector* we call an n-tuple of real numbers $X = (x_1, ..., x_n)$. Operations with vectors are defined by the last three columns of Table 5.2. When $n = 1$, vectors are just numbers and those operations are just usual summation and multiplication. Note that sometimes we need the product XY defined similarly using coordinate-wise multiplication $XY = (x_1 y_1, ..., x_n y_n)$. These formal definitions plus the geometric interpretation considered next is all you need to know about vectors.

(3) Geometric interpretation. We live in the space of three-dimensional vectors. All our intuition coming from day-to-day geometrical experience carries over to the n-dimensional case. Let, for simplicity, $n = 2$. A vector $X = (x_1, x_2)$ is shown on the plane as an arrow starting at the origin and ending at the point (x_1, x_2). The sum $X + Y = (x_1 + y_1, ..., x_2 + y_2)$ is found using the *parallelogram rule* in **Figure 5.1**. The rule itself comes from physics: if two forces are applied to a point, their resultant force is found by the parallelogram rule.
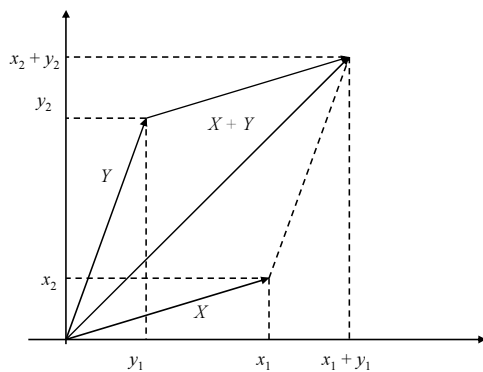

Figure 5.1 Summing vectors using the parallelogram rule

Scaling X by a positive number a means lengthening it in case $a > 1$ and shortening in case $a < 1$. Scaling by a negative number means, additionally, reverting the direction of X.
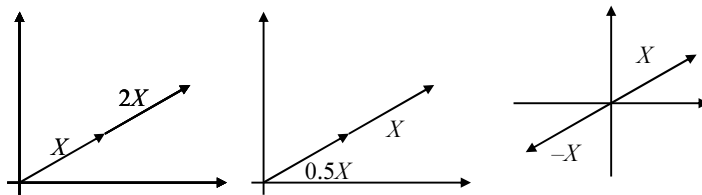

Figure 5.2 Scaling a vector

To obtain a linear combination $aX + bY$ you first scale X and Y and then add the results.

Unit 5.4   Linearity (portfolio, portfolio value, universal statements, existence statements, homogeneity of degree 1, additivity)

Linearity is one of those many properties in which both the statement of the problem and its solution are mathematical.

Problem statement. What is the relation between $E(aX + bY)$ and EX, EY?

Answer. For any random variables X, Y and any numbers a, b one has
$E(aX + bY) = aEX + bEY$.

Proof. This is one of those straightforward proofs when knowing the definitions and starting with the left-hand side is enough to arrive at the result. Using the definitions in Table 5.2 the mean of the linear combination is

$$E(aX + bY) = (ax_1 + by_1)p_1 + ... + (ax_n + by_n)p_n \qquad \text{(multiplying out)}$$
$$= ax_1p_1 + by_1p_1 + ... + ax_np_n + ... + by_np_n \quad \text{(grouping by variables)}$$
$$= (ax_1p_1 + ... + ax_np_n) + (by_1p_1 + ... + by_np_n)$$
$$= aEX + bEY.$$

Chewing and digesting

Going back. A *portfolio* contains n_1 shares of stock 1 whose price is S_1 and n_2 shares of stock 2 whose price is S_2. Stock prices fluctuate and are random variables. Numbers of shares are assumed fixed and are deterministic. What is the expected value of the portfolio?

Answer. The *portfolio value* is its market price $V = n_1 S_1 + n_2 S_2$. Since this is a linear combination, the expected value is $EV = n_1 ES_1 + n_2 ES_2$.

Going sideways. **(1)** Rewrite the proof of linearity of means using summation signs.

(2) Do you think $E(X^2)$ is the same as $(EX)^2$? Hint: It should be clear that $X^2 = (x_1^2, ..., x_n^2)$ and $EX^2 = x_1^2 p_1 + ... + x_n^2 p_n$, see Unit 5.3.

(3) While I don't give an answer for the previous question, I want to use it to discuss the difference between *universal statements* (which should be true for a class of objects) and *existence statements* (which assert existence of objects with special properties). If you say: Yes, I think that $E(X^2) = (EX)^2$, it should be a universal statement: for ANY random variable X one has $E(X^2) = (EX)^2$. If, on the other hand, you say: No, I don't think so, then it should be an existence statement: I am producing X such that $E(X^2) \neq (EX)^2$ (whereas

for other variables the equation may be true). Pay attention to the order of words when you use "for every" and "there exists". In the next two propositions the words are the same, just their order is different:

(i) In the city of X, for every man M there is a woman W such that W is a girlfriend of M.

(ii) In the city of X, there is a woman W such that W is a girlfriend of M for every man M.

Going forward. **(1)** Letting in the linearity property $b = 0$ we get $E(aX) = aEX$. This property is called *homogeneity of degree* 1 (you can pull the constant out of the expected value sign).

(2) Letting $a = b = 1$ in the linearity property we get $E(X + Y) = EX + EY$. This is called *additivity*.

(3) Use induction to prove the generalization to the case of a linear combination of n variables:

$$E(a_1 X_1 + ... + a_n X_n) = a_1 EX_1 + ... + a_n EX_n$$

(4) Expected value of a constant. The mean of a constant is that constant, $Ec = cp_1 + ... + cp_n = c(p_1 + ... + p_n) = 1$, because a constant doesn't change, rain or shine.

Unit 5.5   Independence and its consequences (independent and dependent variables, multiplicativity of means)

Background. You have to recall the definition of independence of two events $P(A \cap B) = P(A)P(B)$ and realize that a random variable is a much more complex object than an event. Visually, a random variable is a 1-D table with values and probabilities. A pair of random variables is represented by a 2-D table similar to **Table 4.8**. Since we want to work with variables whose probability distributions are not necessarily the same, a table of type **Table 5.2** is insufficient. We deal with two random variables separately represented by the following tables

Table 5.3 Separate probability tables

Values of X	Probabilities		Values of Y	Probabilities
x_1	p_1^X		y_1	p_1^Y
…	…		…	…
x_n	p_n^X		y_m	p_m^Y

In general, n is different from m. The probabilities are also different, that's why they are provided with superscripts. If you remember, in this situation we cannot directly define the sum $X + Y$ and the product XY. This is why we need to work with the pair (X, Y) which has values (x_i, y_j), $i = 1, ..., n$, $j = 1, ..., m$ (you can think about throwing a coin and die at the same time; any value on the coin can occur in combination with any value on the die). The probabilities of these values are denoted p_{ij}, $i = 1, ..., n$, $j = 1, ..., m$. Now we arrange a 2-D table putting a pair of values (x_i, y_j) and corresponding probability p_{ij} in the same cell:

Table 5.4 Joint values and probabilities

	y_1	…	y_m	Right margin

x_1	(x_1, y_1), p_{11}	$\cdots$	(x_1, y_m), p_{1m}	p_1^X
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
x_n	(x_n, y_1), p_{n1}	$\cdots$	(x_n, y_m), p_{nm}	p_n^X
Lower margin	p_1^Y	$\cdots$	p_m^Y	1

Exercise 5.1 One and the same random variable can have many different (equivalent) realizations. The variable X is realized in **Table 5.3** on the sample space $\{x_1,...,x_n\}$. What is its realization in terms of **Table 5.4** (what are the sample space, probability distribution and values)?

Exercise 5.2 You can refresh your memory by rewriting in this notation the identities following **Table 4.8**.

Intuitive definition. We say that X and Y are *independent* if what happens to X does not influence Y in any way.

Formal definition. We say that X and Y are *independent* if the events independence condition is satisfied for every pair of their values:

$$p_{ij} = p_i^X p_j^Y, \ i = 1,...,n, \ j = 1,...,m \tag{5.3}$$

Comments. **(1)** This is an opportune moment to discuss the relation between universal statements and existence statements. Consider a universal statement: all students in a given class are younger than 30. Its opposite is: there are students in a given class who are at least 30 years old, which is an existence statement. This is a general rule of logic: rejecting a universal statement you obtain an existence one, and vice versa.

(2) Rejecting the definition of independence, we obtain the definition of dependence, and it should be an existence statement. Thus, we say that X and Y are *dependent* if at least one of the equations (5.3) is not true.

Definition. The sum $X+Y$ and the product XY are defined by the tables

Table 5.5 General definitions of the sum and product

	Definition of $X+Y$				Definition of XY		
	y_1	$\cdots$	y_m		y_1	$\cdots$	y_m
x_1	$x_1 + y_1$	$\cdots$	$x_1 + y_m$	x_1	$x_1 y_1$	$\cdots$	$x_1 y_m$
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
x_n	$x_n + y_1$	$\cdots$	$x_n + y_m$	x_n	$x_n y_1$	$\cdots$	$x_n y_m$

Exercise 5.3 How is the definition related to the earlier one?

Multiplicativity of expected values. If X, Y are independent, then their mean is *multiplicative*: $EXY = (EX)(EY)$.

Proof. Earlier you were asked to rewrite the proof of linearity of means using summation signs. Here we need a special case of that property:

$$\sum_{j=1}^{m} a z_j = a \sum_{j=1}^{m} z_j .$$

Using **Table 5.5** and probabilities from **Table 5.4** we have

$$EXY = \sum_{i=1}^{n} \sum_{j=1}^{m} x_i y_j p_{ij} \qquad \text{(using independence assumption)}$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{m} x_i y_j p_i^X p_j^Y \qquad (a = x_i p_i^X \text{ is a constant for the inner sum})$$

$$= \sum_{i=1}^{n} x_i p_i^X \sum_{j=1}^{m} y_j p_j^Y = (EX)(EY).$$

Comments. **(1)** If you don't understand this, take $n = m = 2$, write out everything without summation signs and then write out equivalents with summation signs.

(2) Multiplicativity is not a linear property. All operators in math that are linear may have nonlinear properties only under special conditions. Independence is such a special condition. Recapitulating, $E(aX + bY) = aEX + bEY$ is true for ANY X, Y and $EXY = (EX)(EY)$ holds ONLY under independence.

(3) The complete notation for the fact that X takes the value x_i with probability p_i^X is $P(X = x_i) = p_i^X$. Similarly, the complete expression for the probabilities from **Table 5.4** is $P(X = x_i, Y = y_j) = p_{ij}$. This formalism is necessary to understand some equations in NCT.

(4) Two other characteristics of random variables, variance and covariance, are defined in terms of means. Therefore we shall be using linearity of means very often.

Unit 5.6 Covariance (linearity, alternative expression, uncorrelated variables, sufficient and necessary conditions, symmetry of covariances)

Definition. For two random variables X, Y their *covariance* is defined by
$\text{cov}(X, Y) = E(X - EX)(Y - EY)$.

Comments. **(1)** Covariance has two complex arguments, whereas its value is simple – just a number.

(2) $X - EX$ is the deviation of X from its mean; $Y - EY$ is the deviation of Y from its mean. Therefore $\text{cov}(X, Y)$ is the expected value of the product of these two deviations. Explain all formulas for yourself in this way.

(3) The book says that covariance is a measure of relationship between two variables. This is true with some caveats. We keep covariance around mainly for its algebraic properties.

Property 1. Linearity. Covariance is *linear* in the first argument when the second argument is fixed: for any random variables X, Y, Z and numbers a, b one has

$$\text{cov}(aX + bY, Z) = a \, \text{cov}(X, Z) + b \, \text{cov}(Y, Z). \tag{5.4}$$

Proof. We start by writing out the left side of (5.4):

$$\text{cov}(aX + bY, Z)$$
$$= E[(aX + bY) - E(aX + bY)](Z - EZ) \qquad \text{(using linearity of means)}$$
$$= E(aX + bY - aEX - bEY)(Z - EZ) \qquad \text{(collecting similar terms)}$$
$$= E[a(X - EX) + b(Y - EY)](Z - EZ) \qquad \text{(multiplying out)}$$
$$= E[a(X - EX)(Z - EZ) + b(Y - EY)(Z - EZ)] \quad \text{(using linearity of means)}$$
$$= aE(X - EX)(Z - EZ) + bE(Y - EY)(Z - EZ) \quad \text{(by definition of covariance)}$$
$$= a\,\text{cov}(X, Z) + b\,\text{cov}(bY, Z).$$

Exercise 5.4 Covariance is also linear in the second argument when the first argument is fixed. Write out and prove this property. You can notice the importance of using parentheses and brackets.

> Learning longer proofs will develop your logic and intuition to the extent that you'll be able to produce results like this on demand.

Property 2. *Alternative expression*: $\text{cov}(X, Y) = EXY - (EX)(EY)$.

Proof.

$$\text{cov}(X, Y)$$
$$= E(X - EX)(Y - EY) \qquad \text{(multiplying out)}$$
$$= E[XY - XEY - (EX)Y + (EX)(EY)] \qquad (EX, EY \text{ are constants; use linearity)}$$
$$= EXY - (EX)(EY) - (EX)(EY) + (EX)(EY)$$
$$= EXY - (EX)(EY).$$

Definition. Random variables X, Y are *uncorrelated* if $\text{cov}(X, Y) = 0$. Uncorrelatedness is close to independence, so the intuition is the same: one variable does not influence the other. You can also say that there is no statistical relationship between uncorrelated variables.

Property 3. Independent variables are uncorrelated: if X, Y are independent, then $\text{cov}(X, Y) = 0$.

Proof. By the alternative expression of covariance and multiplicativity of means for independent variables $\text{cov}(X, Y) = EXY - (EX)(EY) = 0$.

Comments. **(1)** This is a good place to discuss necessary and sufficient conditions. We say that condition A is *sufficient* for condition B if A implies B: A $\Rightarrow$ B. When this is true, we also say that B is *necessary* for A. Visually, the set of all objects satisfying A is a subset of all objects satisfying B when A is sufficient for B. For example, the statement on relationship between independence and uncorrelatedness is visualized like this:
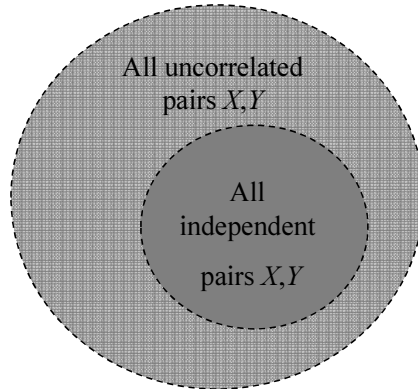
Figure 5.3 Independent versus uncorrelated pairs of variables

(2) When A is both necessary and sufficient for B, we say that A is *equivalent* to B and write A $\Leftrightarrow$ B. An interesting question is whether independence is equivalent to uncorrelatedness or, put it differently, if there exist uncorrelated pairs which are not independent. The answer is Yes. Take any variable X such that $EX = EX^3 = 0$ and put $Y = X^2$. This Y is not independent of X because knowledge of the latter implies knowledge of the former. However, they are uncorrelated:

$$\text{cov}(X,Y) = EXY - (EX)(EY) = EX^3 - (EX)(EY) = 0.$$

(3) Here we show that the condition $EX = EX^3 = 0$ is satisfied for any random variable that is symmetric around zero. The symmetry definition implies that any negative value x_i^- can be written as $x_i^- = -x_i^+$ where x_i^+ is a positive value of the same variable and the probabilities are the same: $P(X = x_i^-) = P(X = x_i^+)$. Therefore

$$
\begin{aligned}
EX &= \sum x_i p_i & \text{(sorting out positive and negative values)} \\
&= \sum_{x_i^- < 0} x_i^- p_i + \sum_{x_i^+ > 0} x_i^+ p_i & \text{(replacing negative values and their probabilities)} \\
&= \sum_{x_i^+ > 0} [(-x_i^+) + x_i^+] P(X = x_i^+) & \text{(the two terms cancel out)} \\
&= 0.
\end{aligned}
$$

The proof of $EX^3 = 0$ is similar.

Property 4. Correlation with a constant. Any random variable is uncorrelated with any constant: $\text{cov}(X,c) = E(X - EX)(c - Ec) = 0$.

Property 5. Symmetry. Covariance is a *symmetric function* of its arguments: $\text{cov}(X,Y) = \text{cov}(Y,X)$.

Unit 5.7   Variance (interaction term, variance of a sum, homogeneity of degree 2, additivity of variance)

Definition. For a random variable X, its variance is defined by $V(X) = E(X - EX)^2$.

Comments. **(1)** Variance measures the spread of X around its mean.

(2) For linear operations, like the mean EX, it is common to write the argument without parentheses, unless it is complex, as in $E(aX + bY)$. For nonlinear operations, of type $V(X)$, the argument must be put in parentheses.

Property 1. Relation to covariance: $V(X) = \text{cov}(X, X)$.

Property 2. Variance of a linear combination. For any random variables X, Y and numbers a, b one has

$$V(aX + bY) = a^2 V(X) + 2ab \, \text{cov}(X, Y) + b^2 V(Y). \tag{5.5}$$

The term $2ab \, \text{cov}(X, Y)$ in (5.5) is called an *interaction term*.

Proof. Use Property 1:

$$V(aX + bY) = \text{cov}(aX + bY, aX + bY)$$

(using linearity with respect to the first variable)

$$= a \, \text{cov}(X, aX + bY) + b \, \text{cov}(Y, aX + bY)$$

(using linearity with respect to the second variable)

$$= a^2 \, \text{cov}(X, X) + ab \, \text{cov}(X, Y) + ab \, \text{cov}(Y, X) + b^2 \, \text{cov}(Y, Y)$$

(using symmetry of covariance and collecting similar terms)

$$= a^2 V(X) + 2ab \, \text{cov}(X, Y) + b^2 V(Y).$$

Exercise 5.5 (1) Prove the property directly, without appealing to covariance.

(2) How do you obtain from this property two special cases:

(i) $V(X + Y) = V(X) + 2 \, \text{cov}(X, Y) + V(Y)$ (*variance of a sum*)

and

(ii) $V(aX) = a^2 V(X)$ (*homogeneity of degree 2*)?

(3) What do you think is larger: $V(X + Y)$ or $V(X - Y)$?

(4) If we add a constant to a variable, does its variance change?

Property 3. Variance of a constant: $V(c) = E(c - Ec)^2 = 0$.

Property 4. Nonnegativity: $V(X) \geq 0$ for any X.

Proof. By definition,

$$V(X) = E(X - EX)^2 = p_1(x_1 - EX)^2 + \ldots + p_n(x_n - EX)^2. \tag{5.6}$$

This is a weighted sum of squared deviations, each of which is nonnegative, and the weights are positive, so the whole thing is nonnegative.

Property 5. Characterization of all variables with vanishing variance. We want to know exactly which variables have zero variance. We know that constants have this property, so $X = c$ is sufficient for $V(X) = 0$. Conversely, suppose that $V(X) = 0$. Since all probabilities are positive, from (5.6) we see that $x_1 - EX = \ldots = x_n - EX = 0$. This implies that all values are equal to the mean, which is a constant. Conclusion: $V(X) = 0$ if and only if $X = c$.

Property 6. *Additivity of variance*: if X, Y are independent, then $V(X + Y) = V(X) + V(Y)$ (the interaction terms disappears).

Since variance is a nonlinear operator, it is additive only under special circumstances, when the variables are uncorrelated or, in particular, when they are independent.

Unit 5.8 Standard deviation (absolute value, homogeneity, Cauchy-Schwarz inequality)

Definition. For a random variable X, the quantity $\sigma(X) = \sqrt{V(X)}$ is called its *standard deviation.*

In general, there are two square roots of a positive number, one positive and the other negative. The positive one is called an *arithmetic square root*. An arithmetic root is applied here to $V(X) \geq 0$, so standard deviation is always nonnegative.

Absolute values. An *absolute value* of a real number is defined by

$$| a | = \begin{cases} a, & \text{if } a \text{ is nonnegative} \\ -a, & \text{if } a \text{ is negative} \end{cases}$$

(5.7)

It is introduced to measure the distance from point a to the origin. For example, dist(3,0) = |3| = 3 and dist(–3,0) = |–3| = 3. More generally, for any points a, b on the real line the distance between them is given by $\text{dist}(a, b) = | a - b |$.

The distance interpretation of the absolute value makes it clear that

(i) the condition $a \in [-b, b]$ is equivalent to $| a | \leq b$

and

(ii) the condition $a \notin [-b, b]$ is equivalent to $| a | > b$.

By squaring both sides in (5.7) we get $| a |^2 = a^2$. Application of the arithmetic square root gives $| a | = \sqrt{a^2}$. This is the equation we need right now.

Property 1. Standard deviation is *homogeneous of degree* 1:
$\sigma(aX) = \sqrt{V(aX)} = \sqrt{a^2 V(X)} = | a | \sigma(X)$.

Property 2. *Cauchy-Schwarz inequality*. **(1)** For any random variables X, Y one has $| \text{cov}(X, Y) | \leq \sigma(X) \sigma(Y)$.

(2) If the inequality sign turns into equality, $| \text{cov}(X, Y) | = \sigma(X) \sigma(Y)$, then Y is a linear function of X: $Y = aX + b$.

Proof. **(1)** If at least one of the variables is constant, both sides of the inequality are 0 and there is nothing to prove. To exclude the trivial case, let X, Y be nonconstant. Consider a real-valued function of a real number t defined by $f(t) = V(tX + Y)$. Here we have variance of a linear combination: $f(t) = t^2 V(X) + 2t \, \text{cov}(X, Y) + V(Y)$.

We see that $f(t)$ is a parabola with branches looking upward (because the senior coefficient $V(X)$ is positive). By nonnegativity of variance, $f(t) \geq 0$ and the parabola lies above the horizontal axis in the (f, t) plane. Hence, the quadratic equation $f(t) = 0$ may have at most

one real root. This means that the discriminant of the equation is non-positive: $D = [\text{cov}(X,Y)]^2 - V(X)V(Y) \le 0$. Applying square roots to both sides of $[\text{cov}(X,Y)]^2 \le V(X)V(Y)$ we finish the proof of the first part.

(2) In case of the equality sign the discriminant is 0. Therefore the parabola touches the horizontal axis where $f(t) = V(tX + Y) = 0$. But we know that this implies $tX + Y = c$ which is just another way of writing $Y = aX + b$.

> Do you think this proof is tricky? During the long history of development of mathematics mathematicians have invented many tricks, small and large. No matter how smart you are, you cannot reinvent all of them. If you want to learn a lot of math, you'll have to study the existing tricks. By the way, the definition of what is tricky and what is not is strictly personal and time-dependent.

Unit 5.9 Correlation coefficient (unit-free property, positively and negatively correlated variables, uncorrelated variables, perfect correlation)

Definition. Suppose random variables X, Y are not constant. Then their standard deviations are not zero and we can define

$$\rho(X,Y) = \frac{\text{cov}(X,Y)}{\sigma(X)\sigma(Y)}$$

which is called a *correlation coefficient* between X and Y.

Property 1. *Range of the correlation coefficient*: for any X, Y one has $-1 \le \rho(X,Y) \le 1$.

Proof. By the properties of absolute values the Cauchy-Schwarz inequality can be written as

$$-\sigma(X)\sigma(Y) \le \text{cov}(X,Y) \le \sigma(X)\sigma(Y).$$

It remains to divide this inequality by $\sigma(X)\sigma(Y)$.

Property 2. **Interpretation of extreme cases**. **(1)** If $\rho(X,Y) = 1$, then $Y = aX + b$ with $a > 0$.

(2) If $\rho(X,Y) = -1$, then $Y = aX + b$ with $a < 0$.

Proof. **(1)** $\rho(X,Y) = 1$ implies

$$\text{cov}(X,Y) = \sigma(X)\sigma(Y) \tag{5.8}$$

which, in turn, implies that Y is a linear function of X : $Y = aX + b$. Further, we can establish the sign of the number a. By the properties of variance and covariance

$$\text{cov}(X,Y) = \text{cov}(X, aX + b) = a\,\text{cov}(X,X) + \text{cov}(X,b) = aV(X),$$
$$\sigma(Y) = \sigma(aX + b) = \sigma(aX) = |a|\sigma(X).$$

Plugging this in (5.8) we get $aV(X) = |a|\sigma^2(X)$ and see that a is positive.

The proof of **(2)** is left as an exercise.

Property 3. Suppose we want to measure correlation between weight W and height H of people. The measurements are either in kilos and centimeters W_k, H_c or in pounds and feet

W_p, H_f. The correlation coefficient is *unit-free* in the sense that it does not depend on which units are used: $\rho(W_k, H_c) = \rho(W_p, H_f)$.

Proof. One measurement is proportional to another, $W_k = aW_p$, $H_c = bH_f$ with some positive constants a, b. By homogeneity

$$\rho(W_k, H_c) = \frac{\text{cov}(W_k, H_c)}{\sigma(W_k)\sigma(H_c)} = \frac{\text{cov}(aW_p, bH_f)}{\sigma(aW_p)\sigma(bH_f)} = \frac{ab\,\text{cov}(W_p, H_f)}{ab\sigma(W_p)\sigma(H_f)} = \rho(W_p, H_f).$$

Definition. Because the range of $\rho(X, Y)$ is $[-1, 1]$, we can define the angle $\alpha_{X,Y}$ between X and Y by $\cos\alpha_{X,Y} = \rho(X, Y)$, $0 \leq \alpha_{X,Y} \leq \pi$.

Table 5.6 Geometric interpretation of the correlation coefficient

X and Y are *positively correlated*: $0 < \rho \leq 1$	$0 \leq \alpha_{X,Y} < \dfrac{\pi}{2}$ (acute angle)
X and Y are *negatively correlated*: $-1 \leq \rho < 0$	$\dfrac{\pi}{2} < \alpha_{X,Y} \leq \pi$ (obtuse angle)
X and Y are *uncorrelated*: $\rho = 0$	$\alpha_{X,Y} = \dfrac{\pi}{2}$ (right angle)
X and Y are *perfectly positively correlated*: $\rho = 1$	$Y = aX + b$ with $a > 0$
X and Y are *perfectly negatively correlated*: $\rho = -1$	$Y = aX + b$ with $a < 0$

Unit 5.10 Standardization (standardized version)

Questions. The **motivation** here is purely mathematical. **(1)** What can be done to a random variable X to obtain a new random variable with mean zero?

(2) What can be done to X to obtain a new variable with unit variance (and standard deviation)?

Digression on linear transformations. Let us look at what happens to the distribution of X when a linear transformation $Y = aX + b$ is applied to it. From Table 5.2 it is clear that $Y = aX + b$ takes values $ax_i + b$.

Consider a special case $a = 1$. When a constant b is added to X, all values of X move to the right ($b > 0$) or to the left ($b < 0$) by the same amount. Correspondingly, the histogram moves right or left as a rigid body (the distances between values do not change).

Consider another special case $b = 0$. If X is multiplied by a, its values will be stretched ($|a| > 1$) or contracted towards the origin ($|a| < 1$). In case of stretching its variance increases and in case of contraction it decreases.

In the general case stretching/contraction is combined with the movement along the x-axis.

Answers. **(1)** To obtain a variable with mean zero out of X we subtract its mean: $E(X - EX) = EX - EX = 0$.

(2) To obtain a variable with unit variance out of X we divide it by $\sigma(X)$ (provided it's not zero):

$$V\left(\frac{X}{\sigma(X)}\right) = \text{(by homogeneity)} = \frac{1}{\sigma^2(X)}V(X) = 1.$$

To satisfy both requirements we combine the two transformations:

$$Y = \frac{X - EX}{\sigma(X)}. \tag{5.9}$$

Definition. The transformation of X defined in (5.9) is called *standardization* and Y is called a *standardized version* of X.

Exercise 5.6 Check that (5.9) has mean zero and unit variance.

In NCT the mean, variance and standard deviation are denoted by μ, σ^2 and σ, respectively. Always put a subscript to reflect the variable in question in this notation:
$$\mu_X = EX, \ \sigma_X^2 = V(X), \ \sigma_X = \sqrt{V(X)} = \sigma(X).$$

Unit 5.11   Bernoulli variable (population, sample, sample size, sampling with replacement, ex-post and ex-ante treatment of random variables, identically distributed variables, parent population, binomial variable, number of successes, proportion of successes, standard error)

Intuitive definition. A *Bernoulli variable* is the variable that describes an unfair coin.

Formal definition. A *Bernoulli variable* is defined by the table:

Table 5.7 Bernoulli variable definition

Values of B	Probabilities
1	p
0	$1-p$

where $0 < p < 1$. Often it is convenient to denote $q = 1 - p$.

Property 1. *Mean of the Bernoulli variable*: $EB = 1 \cdot p + 0 \cdot (1-p) = p$.

Property 2. *Variance of the Bernoulli variable*. Using the alternative expression for variance,

$$\begin{aligned} V(B) &= EB^2 - (EB)^2 \\ &= 1^2 p + 0^2(1-p) - p^2 = p - p^2 = p(1-p). \end{aligned}$$

Note that $V(B)$ is a parabola with branches looking downward. It vanishes at $p = 0$ and $p = 1$ and is maximized at $p = 1/2$, which corresponds to a fair coin.

5.11.1   Some facts about sampling not many people will tell you

Here we provide an example of an unfair coin and use it to discuss the notions of population and sample.

Take any book in English. All alphabetic characters in it are categorized as either consonants or vowels. Denote by p the percentage of consonants. We assign 1 to a consonant and 0 to a vowel. The experiment consists in randomly selecting an alphabetic character from the book. Then the outcome is 1 with probability p and 0 with probability q. We have a random

variable described by Table 5.7. In this context, the population is the set of consonants and vowels, along with their percentages p and q. This description fits the definition from Section 1.2 of NCT: a *population* is a complete set of all items that interest the investigator.

Another definition from the same section says: a *sample* is an observed subset of population values with *sample size* given by n. It is important to imagine the sampling process step by step. To obtain a sample in our example we can randomly select n characters and write the sample as

$$X_1, ..., X_n$$

where some X's are unities and others are zeros. The problem is that "random selection" can be understood in different ways. We consider the most important two.

Case I. We randomly select X_1 from the whole book. Suppose it's the first character in the table of contents, which is "C" in NCT. Then $X_1 = 1$. Then we randomly select X_2, again from the whole book. It may well be the same "C" in the table of contents, in which case $X_2 = 1$. We continue like this following two rules: (1) each selection is random and (2) the selection is made from the whole book. The second condition defines what is called a *sampling with replacement*: to restore the population, selected elements are returned to it. After we have obtained a sample of size n, we can form a sum of the observed values

$$S_n = X_1 + ... + X_n \tag{5.10}$$

and a sample mean

$$\overline{X} = \frac{X_1 + ... + X_n}{n} = \frac{S_n}{n}. \tag{5.11}$$

After we obtain an actual sample, there will be nothing random. S_n, being a sum of unities and zeros, will be some integer from the range $[0, n]$ and $\overline{X}$ will be some fraction from $[0,1]$.

> The subtlety of the theoretical argument in statistics is that $X_1, ..., X_n$ are treated not as realized (*ex-post*) values but as theoretical (*ex-ante*) random variables that are associated with the drawings from the population.

Let us rewind the movie and think about the selection process again. The first time you don't actually select a character, you just THINK about randomly selecting it from the whole book. The associated random variable is denoted X_1 and, clearly, is described by the same table as B. Similarly, you think about what possibilities you would face in all other cases and come to the conclusion that all drawings are described by the same table, just the names of the variables will be different.

Definition. Random variables that have the same distributions (that is, values plus probabilities) are called *identically distributed*.

Comments. **(1)** Since values plus probabilities is all one needs to find the mean and variance of a random variable, identically distributed variables have identical means and variances.

(2) Instead of saying " $X_1, ..., X_n$ are identically distributed" we can equivalently say " $X_1, ..., X_n$ are drawn from the same population". In this case the population we draw the sample from is called a *parent population*.

(3) By construction, in Case I the variables are independent. The common abbreviation for "independent identically distributed" is i.i.d.

(4) In the sum (5.10) the variables $X_1,...,X_n$ may take independently values 0 and 1. If all are zeros, S_n is zero. If all are unities, S_n will be equal to n. In general, S_n may take any integer value from 0 to n, inclusive. To emphasize that we deal with variables, in the equation

$$S_{n \atop 0,...,n} = X_1 + ... + X_n \atop {\scriptstyle 0,1} \qquad {\scriptstyle 0,1}$$

below the variables I write their possible values. Now the next definition must be clear.

Definition. A *binomial variable* is a sum of n independent identically distributed Bernoulli variables.

Case II. Think about an opinion poll on a simple question requiring a simple answer: Yes or No. Suppose everybody has a definite opinion on the matter, and denote p the percentage of voters who have an affirmative answer. A team is sent out to conduct a survey. They record 1 for Yes and 0 for No. The population will be again Bernoulli but there are problems with the sample. It doesn't make sense to ask the same person twice, so the elements in the sample cannot be repeated. In essence, after having randomly selected the first individual, you cross out that individual from the population before selecting the next one. This implies two things: the observations are dependent and their distributions are not identical.

The general statistical principle is that removing one element from a large population does not change its characteristics very much. So in the case under consideration, the observations can be assumed approximately i.i.d.

Conclusion. If $X_1,...,X_n$ are drawn from the same parent population and μ_X and σ_X denote the mean and standard deviation of the parent population, then

$$EX_1 = ... = EX_n = \mu_X, \ \sigma(X_1) = ... = \sigma(X_n) = \sigma_X. \tag{5.12}$$

5.11.2   An obvious and handy result

It is possible to calculate means and variances of S_n and $\overline{X}$ by deriving their distributions and applying the definitions of a mean and variance directly (see **Exercise 5.8**). But there is a much easier way to do that, which should be obvious, if not now then at least after reading the proof below. It is as handy as it is indispensable in a number of applications.

The sum $S_n = X_1 + ... + X_n$ and the sample mean $\overline{X}$ can be used when X_i are drawn from an arbitrary population. In the context of drawing from a Bernoulli population S_n is called a *number of successes* (because only unities contribute to it). The sample mean $\overline{X}$ in this case is called a *proportion of successes* and denoted $\hat{p}$.

Exercise 5.7 The mean and variance of S_n and $\overline{X}$. (1) If $X_1,...,X_n$ are i.i.d., then

$$ES_n = n\mu_X, \ V(S_n) = n\sigma_X^2, \ E\overline{X} = \mu_X, \ V(\overline{X}) = \frac{\sigma_X^2}{n}. \tag{5.13}$$

(2) If, additionally, S_n is a binomial variable and p_X denotes the parameter of the parent Bernoulli population, then (5.13) becomes

$$ES_n = np_X, \ V(S_n) = np_X(1-p_X), \ E\hat{p} = p_X, \ V(\hat{p}) = \frac{p_X(1-p_X)}{n}. \qquad (5.14)$$

Proof. **(1)** Here our derivation of algebraic properties of means and variances starts paying back.

$$\begin{aligned} ES_n &= E(X_1 + ... + X_n) \quad \text{(by additivity)} \\ &= EX_1 + ... + EX_n \quad (X_1, ..., X_n \text{ are identically distributed}) \\ &= nEX = n\mu_X. \end{aligned}$$

For variance we use also independence. Due to independence, all interaction terms vanish and variance is additive.

$$\begin{aligned} V(S_n) &= V(X_1 + ... + X_n) \quad \text{(by additivity)} \\ &= V(X_1) + ... + V(X_n) \quad (X_1, ..., X_n \text{ are identically distributed}) \\ &= nV(X) = n\sigma_X^2. \end{aligned}$$

The results for the sample mean follow by homogeneity:

$$E\overline{X} = E\frac{S_n}{n} = \frac{1}{n} ES_n = \frac{1}{n} n\mu_X = \mu_X,$$

$$V(\overline{X}) = V\left(\frac{S_n}{n}\right) = \frac{1}{n^2} V(S_n) = \frac{1}{n^2} n\sigma_X^2 = \frac{\sigma_X^2}{n}.$$

(2) In case of the binomial variable just replace μ_X by p_X and σ_X^2 by $p_X(1-p_X)$.

From this exercise it follows that $\sigma_{\overline{X}} = \frac{\sigma_X}{\sqrt{n}}$. The quantity on the left is called a *standard error* of $\overline{X}$.

An important consequence is that $V(\overline{X}) \to \infty$ as $n \to \infty$ (as the sample increases, the estimation precision improves and the values of the sample mean become more and more concentrated around the population mean; is this result intuitive?) This is a general statistical fact: increasing the sample is always good.

Unit 5.12 Distribution of the binomial variable

Case $n = 3$. When you need to recall what a binomial variable is, you start with "it is a sum" and then disclose the rest. Let us derive the distribution of S_3. The idea is to list all possible combinations of values of X_1, X_2, X_3, find the corresponding values of S_3 and then calculate their probabilities. This is done in two steps. Remember that the objective is to get the "values plus probabilities" table.

> In mathematical exposition ideas sometimes are not stated. The presumption is that you get the idea after reading the proof. In your mind, ideas should always go first.

Step 1.

Table 5.8 Table of possible outcomes and probabilities

Row	Combinations of values of X_1, X_2, X_3	Values of S_3	Probabilities
1	(0,0,0)	0	q^3

2	(1,0,0)	1	pq^2
3	(0,1,0)	1	pq^2
4	(0,0,1)	1	pq^2
5	(1,1,0)	2	p^2q
6	(0,1,1)	2	p^2q
7	(1,0,1)	2	p^2q
8	(1,1,1)	3	p^3

Here we used the multiplication rule. For example, the probability pq^2 in the second row is obtained like this:

$$P(X_1 = 1, X_2 = 0, X_3 = 0) = P(X_1 = 1)P(X_2 = 0)P(X_3 = 0) = pq^2.$$

Step 2. The next step is to collect equal values of S_3. This is done using the additivity rule. For example, the events (1,0,0), (0,1,0) and (0,0,1) in rows 2-4 are mutually exclusive and therefore

$$P(S_3 = 1) = P((1,0,0) \cup (0,1,0) \cup (0,0,1))$$
$$= P((1,0,0)) + P((0,1,0)) + P((0,0,1)) = 3pq^2.$$

Thus the result is

Table 5.9 Distribution of S_3

Values of S_3	Probabilities
0	q^3
1	$3pq^2$
2	$3p^2q$
3	p^3

Exercise 5.8 Using this table find $ES_3, V(S_3)$.

Case of general n. The first thing to understand is that S_n takes all integer values x from 0 to n, inclusive, and for each of these values we need to figure out the probability $P(S_n = x)$ (that's what Table 5.9 is about in case $n = 3$).

Step 1. The equation $S_n = x$ tells us that in a sequence $(X_1,...,X_n)$ of zeros and unities there are x unities and $n-x$ zeros. The probability of ONE such sequence is $p^x q^{n-x}$ by the multiplication rule.

Step 2. We have to find out how many such sequences exist for the given x. This question can be rephrased as follows: in how many different ways x places (where the unities will be put) can be chosen out of n places? Since the order doesn't matter, the answer is: in C_x^n ways. Because each of these ways has the same probability $p^x q^{n-x}$ and they are mutually exclusive, by the additivity rule

$$P(S_n = x) = C_x^n p^x q^{n-x}, \ x = 0,...,n. \tag{5.15}$$

Unit 5.13 Distribution function (cumulative distribution function, monotonicity, interval formula, cumulative probabilities)

The full name from NCT is a *cumulative distribution function* but I am going to stick to the short name (used in more advanced texts). This is one of the topics most students don't get on the first attempt.

Example 5.2. Electricity consumption sharply increases when everybody starts using air conditioners, and this happens when temperature exceeds 20°. The utility company would like to know the likelihood of a jump in electricity consumption tomorrow.

(i) Consider the probability $P(T \leq 15)$ that T will not exceed 15°. How does it relate to the probability $P(T \leq 20)$?

(ii) Suppose in the expression $P(T \leq t)$ the real number t increases to $+\infty$. What happens to the probability?

(iii) Now think about t going to $-\infty$. Then what happens to $P(T \leq t)$?

Definition. Let X be a random variable and x a real number. The *distribution function* F_X of X is defined by

$$F_X(x) = P(X \leq x).$$

Answers

(i) $F_X(x)$ is the probability of the event $\{X \leq x\}$, so the value $F_X(x)$ belongs to [0,1]. As the event becomes wider, the probability increases. This property is called *monotonicity* and is formally written as follows: if $x_1 \leq x_2$, then $\{X \leq x_1\} \subset \{X \leq x_2\}$ and $F_X(x_1) \leq F_X(x_2)$.

(ii) As x goes to $+\infty$, the event $\{X \leq x\}$ approaches a sure event and $F_X(x)$ tends to 1.

(iii) As x goes to $-\infty$, the event $\{X \leq x\}$ approaches an impossible event and $F_X(x)$ tends to 0.

(iv) From (i)-(iii) the following graph of the distribution function emerges:
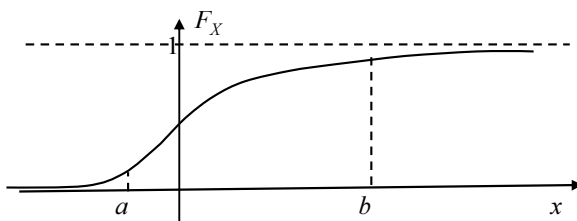
Figure 5.4 Distribution function shape

Most values of X are concentrated where the growth of $F_X(x)$ is the fastest (say, between a and b).

(v) In many applications we are interested in probability of an event that X takes values in an interval $\{a < X \leq b\}$. Such probability can be expressed in terms of the distribution function. Just apply the additivity rule to the set equation $\{-\infty < X \leq b\} = \{-\infty < X \leq a\} \cup \{a < X \leq b\}$ to get $F_X(b) = F_X(a) + P(a < X \leq b)$ and, finally,

$$P(a < X \leq b) = F_X(b) - F_X(a). \tag{5.16}$$

This equation can be called an *interval formula*.

(vi) The definition of $F_X(x)$ equally applies to discrete and continuous variables. However, in case of a discrete random variable the argument x is often restricted to the values taken by the variable. From the working definition of a discrete random variable it should be clear that, if the values are ordered, then

$$F_X(x_1) = P(X \le x_1) = P(X = x_1) = p_1,$$
$$F_X(x_2) = P(X \le x_2) = P(X = x_1) + P(X = x_2) = p_1 + p_2,$$
$$...$$
$$F_X(x_n) = p_1 + ... + p_n.$$

The sums arising here are called *cumulative probabilities*. We summarize our findings in a table:

Table 5.10 Table of probabilities and cumulative probabilities

Values of X	Probabilities	Cumulative probabilities
x_1	p_1	$c_1 = p_1$
x_2	p_2	$c_2 = p_1 + p_2$
…	…	
x_n	p_n	$c_n = p_1 + ... + p_n$

Note that probabilities can be restored from the cumulative probabilities:

$$p_1 = c_1, \ p_2 = c_2 - c_1, \ p_n = c_n - c_{n-1}$$

(vii) In Unit 6.2, Property 2, we shall see that a continuous random variable takes any fixed value with probability zero. Therefore for such variables $F_X(x) = P(X \le x) = P(X < x)$.

Unit 5.14   Poisson distribution (monomials, Taylor decomposition)

The derivation of the Poisson distribution from primitive assumptions would be even more difficult than for the binomial variable. Therefore we introduce it axiomatically and then start looking for real-world applications.

Taylor decomposition. The functions $x^0 = 1$, $x^1 = x$, $x^2, ...$ are called *monomials*. Unlike marsupials, they are considered simple. The idea is to represent other functions as linear combinations of monomials. For example, a polynomial $p(x) = a_0 + a_1 x + ... + a_n x^n$ is such a combination. Linear and quadratic functions are special cases of polynomials. If infinite linear combinations, called *Taylor decompositions*, are allowed, then the set of functions representable using monomials becomes much wider. We are interested in the decomposition of the exponential function

$$e^\lambda = 1 + \frac{1}{1!}\lambda + \frac{1}{2!}\lambda^2 + ... + \frac{1}{n!}\lambda^n + ... = \sum_{x=0}^{\infty} \frac{1}{x!}\lambda^x. \tag{5.17}$$

Here λ is any real number (see calculus texts for more information).

Definition. Take $\lambda > 0$ and divide both sides of (5.17) by e^λ to obtain

$$1 = e^{-\lambda} + \frac{1}{1!}\lambda e^{-\lambda} + \frac{1}{2!}\lambda^2 e^{-\lambda} + ... + \frac{1}{n!}\lambda^n e^{-\lambda} + ... = \sum_{x=0}^{\infty} \frac{1}{x!}\lambda^x e^{-\lambda}.$$

We see that quantities $p_x = \frac{1}{x!}\lambda^x e^{-\lambda}$ satisfy the probability postulate. Hence, we can define a random variable X by

$$P(X = x) = \frac{1}{x!}\lambda^x e^{-\lambda}, \; x = 0, 1, 2,...$$

This variable is called a *Poisson distribution*.

Comments. **(1)** This is a discrete random variable with an infinite number of values. An application should be a variable which can potentially take nonnegative integer values, including very large ones. Normally, a mobile network operator, such as T-Mobile, has a very large number of subscribers. Let X denote the number of all customers of T-Mobile who use their mobile service at a given moment. It is random, it takes nonnegative integer values and the number of customers using their mobile phones simultaneously can be very large. Finally, decisions by customers to use their cell phones can be considered independent, unless there is a disaster.

(2) Compare this definition with what is written in Section 5.6 of NCT. "Assumptions of the Poisson Probability Distribution" from that section is an attempt to show how this distribution is derived from primitive assumptions. Since all the accompanying logic has been cut off, the result is a complete mystery.

Exercise 5.9 Calculate the mean and variance of the Poisson distribution.

Solution. In the definition of the mean the sum is, naturally, extended over all nonnegative integers:

$$EX = \sum_{x=0}^{\infty} xp_x = \sum_{x=0}^{\infty} x\frac{1}{x!}\lambda^x e^{-\lambda}$$

(the term with $x = 0$ is actually 0, so summation starts with $x = 1$)

$$= \sum_{x=1}^{\infty} \frac{1}{(x-1)!}\lambda^x e^{-\lambda}$$

(we pull out λ, to obtain matching $x-1$ in the power and factorial)

$$= \lambda\sum_{x=1}^{\infty} \frac{1}{(x-1)!}\lambda^{x-1} e^{-\lambda}$$

(replace $y = x-1$, to get Poisson probabilities)

$$= \lambda\sum_{y=0}^{\infty} \frac{1}{y!}\lambda^y e^{-\lambda} = \lambda\sum_{y=0}^{\infty} p_y = \lambda.$$

With the view to apply the alternative expression for variance, we calculate first

$$EX^2 = \sum_{x=0}^{\infty} x^2 p_x = \sum_{x=0}^{\infty} x^2 \frac{1}{x!} \lambda^x e^{-\lambda}$$

$$= \sum_{x=1}^{\infty} x \frac{1}{(x-1)!} \lambda^x e^{-\lambda}$$

(this time it is possible to cancel out only one x, so we replace $x = x-1+1$)

$$= \sum_{x=1}^{\infty} (x-1+1) \frac{1}{(x-1)!} \lambda^x e^{-\lambda}$$

$$= \sum_{x=1}^{\infty} (x-1) \frac{1}{(x-1)!} \lambda^x e^{-\lambda} + \sum_{x=1}^{\infty} \frac{1}{(x-1)!} \lambda^x e^{-\lambda}$$

(the idea is again to obtain Poisson probabilities)

$$= \lambda^2 \sum_{x=2}^{\infty} \frac{1}{(x-2)!} \lambda^{x-2} e^{-\lambda} + \lambda \sum_{x=1}^{\infty} \frac{1}{(x-1)!} \lambda^{x-1} e^{-\lambda} = \lambda^2 + \lambda.$$

As a result, $V(X) = EX^2 - (EX)^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$.

Conclusion. The Poisson distribution is special in that its mean and variance are the same, $EX = V(X) = \lambda$. This is another indication as to where it can be applied.

Unit 5.15   Poisson approximation to the binomial distribution

The Poisson distribution is applied directly (see Examples 5.11 and 5.12 in NCT) or indirectly, via the binomial distribution. Before taking up the last point, we look at an example.

Exercise 5.10 An insurance company holds fraudulence insurance policies on 6,000 firms. In any given year the probability that any single policy will result in a claim is 0.001. Find the probability that at least three claims are made in a given year.

Solution. **(a) Using the binomial**.

> On the exam you may not have enough time to calculate the answer. Show that you know the material by introducing the right notation and telling the main idea. Ideas go first!

Put $X = 1$ if a claim is made and $X = 0$ if it is not. This is a Bernoulli variable, $P(X = 1) = 0.001 = p$. Denote S_n the binomial variable. By the additivity rule

$$P(S_n \geq 3) = 1 - P(S_n \leq 2) = 1 - P(S_n = 0) - P(S_n = 1) - P(S_n = 2).$$

From this equation and the general formula $P(S_n = x) = C_x^n p^x q^{n-x}$

$$P(S_n \geq 3) = 1 - 0.999^{6000} - \frac{6000!}{1!5999!} 0.001 \cdot 0.999^{5999} - \frac{6000!}{2!5998!} 0.001^2 \cdot 0.999^{5998}$$

$$= 1 - 0.999^{6000} - 6000 \cdot 0.001 \cdot 0.999^{5999} - 3000 \cdot 5999 \cdot 0.001^2 \cdot 0.999^{5998} \quad (5.18)$$

$$= 1 - 0.999^{6000} - 6 \cdot 0.999^{5999} - 3 \cdot 5.999 \cdot 0.999^{5998}.$$

You will die before you finish raising 0.999 to the power 6000 by hand. The trick is to use the properties of the exponential function and natural logarithm: $a = e^{\ln a}$ and $\ln a^b = b \ln a$ for any positive number a and real number b. In our case

$$0.999^{6000} = e^{\ln 0.999^{6000}} = e^{6000 \ln 0.999} = e^{6000(-0.0010005)} = 0.00247 \,.$$

Eventually we get $P(S_n \geq 3) = 0.93812$.

(b) Using the Poisson distribution. The distinctive feature of (5.18) is that 0.999 raised to a high power may become so small that a simple calculator will identify the result with zero. This small number is multiplied by combinations C_x^n which, when n is very large, may be for a simple calculator the same thing as infinity. However, the product of the two numbers, one small and another large, may be manageable.

Mathematicians are reasonably lazy. When they see such numbers as in (5.18), they prefer to develop a theory than apply brute force. The device for the problem at hand is called a *Poisson approximation to the binomial distribution* and sounds as follows. Suppose p is small and n is large but their product $\lambda = np$ is moderate (the book says $\lambda \leq 7$). Then the probability of the binomial taking the value x can be approximated by the probability of the Poisson taking the same value:

$$P(S_n = x) \approx P(\text{Poisson} = x) = \frac{e^{-\lambda} \lambda^x}{x!} \,.$$

This result gives a reasonably good approximation $P(S_n \geq 3) = 0.938031$.

Unit 5.16   Portfolio analysis (rate of return)

In Unit 5.4 we already defined the portfolio value, and that definition corresponds to what the book has. In fact, the portfolio analysis is a little bit different. To explain the difference, we start with fixing two points of view.

(i) I hold a portfolio of stocks. If I want to sell it, I am interested in knowing its market value. In this situation the numbers of shares in my portfolio, which are constant, and the market prices of stocks, which are random, determine the market value of the portfolio, defined in Unit 5.4. The value of the portfolio is a linear combination of stock prices.

(ii) I have a certain amount of money M^0 to invest. Being a gambler, I am not interested in holding a portfolio. I am thinking about buying a portfolio of stocks now and selling it, say, in a year at price M^1. In this case I am interested in the *rate of return* defined by

$$r = \frac{M^1 - M^0}{M^0} \,.$$

M^0 is considered deterministic (current prices are certain) and M^1 is random (future prices are unpredictable). The rate of return thus is random.

Main result. The rate of return on the portfolio is a linear combination of the rates of return on separate assets.

Proof. As it often happens in economics and finance, this result depends on how one understands the things. The initial amount M^0 is invested in n assets. Denoting M_i^0 the amount invested in asset i, we have $M^0 = \sum_{i=1}^{n} M_i^0$. Denoting $s_i = M_i^0 / M^0$ the share (percentage) of M_i^0 in the total investment M^0, we have

$$M^0 = \sum_{i=1}^{n} s_i M^0, \ M_i^0 = s_i M^0.$$

(5.19)

The shares s_i are deterministic. Denote M_i^1 what becomes of M_i^0 in one year and by $M^1 = \sum_{i=1}^{n} M_i^1$ the value of the total investment at the end of the year. Since different assets change at different rates, generally it is not true that $M_i^1 = s_i M^1$. Denote

$$r_i = \frac{M_i^1 - M_i^0}{M_i^0}$$

(5.20)

the rate of return on asset i. Then using (5.19) and (5.20) we get

$$r = \frac{M^1 - M^0}{M^0} = \frac{\sum(1 + r_i)M_i^0 - \sum M_i^0}{M^0} = \frac{\sum r_i M_i^0}{M^0} = \frac{\sum r_i s_i M^0}{M^0} = \sum s_i r_i.$$

This is the main result. Once you know this equation you can find the mean and variance of the rate of return on the portfolio in terms of shares and rates of return on assets.

Unit 5.17   Questions for repetition

1. List and prove all properties of means.

2. List and prove all properties of covariances.

3. List and prove all properties of variances.

4. List and prove all properties of standard deviations.

5. List and prove all properties of the correlation coefficient including the statistical interpretation from Table 3.3.

6. How and why do you standardize a variable?

7. Define and derive the properties of the Bernoulli variable.

8. Explain the term "independent identically distributed".

9. Define a binomial distribution and derive its mean and variance with explanations.

10. Prove equation (5.15).

11. Define a distribution function, describe its geometric behavior and prove the interval formula.

12. Define the Poisson distribution and describe (without proof) how it is applied to the binomial distribution.

13. Minimum required: Exercises 5.13, 5.23, 5.37, 5.47, 5.67, 5.68, 5.82, 5.91, 5.97.

Chapter 6   Continuous Random Variables and Probability Distributions

Unit 6.1   *Distribution function and density (integral, lower and upper limits of integration, variable of integration, additivity with respect to the domain of integration, linear combination of functions, linearity of integrals, order property)*

Please review the information about distribution functions from Unit 5.13. The definition of the distribution function given there is general and applies equally to discrete and continuous random variables. It has been mentioned that in case of a discrete variable X the argument x of its distribution function is usually restricted to the values of X. Now we are going to look into the specifics of the continuous case.

6.1.1 Properties of integrals

Definition. Let the function f be defined on the segment $[a,b]$. By an *integral*

$$\int_a^b f(x)dx \tag{6.1}$$

we mean the area delimited by the graph of f, the x-axis and two vertical lines: one at a (called a *lower limit of integration*) and another at b (called an *upper limit of integration*).
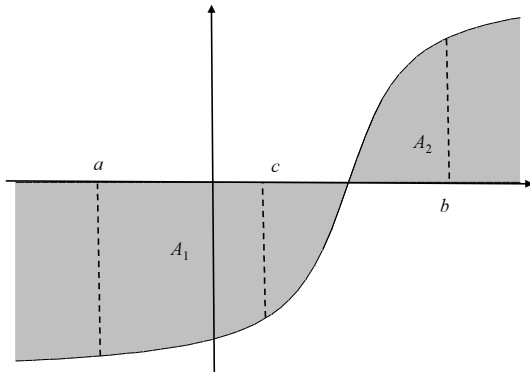
Figure 6.1 Integral as area

Where f is negative, the area is counted with a <u>negative sign</u>. Thus, for the function in Figure 6.1 $\int_a^b f(x)dx = -A_1 + A_2$ where A_1 is the area below the x-axis and above the graph and A_2 is the area above the x-axis and below the graph.

Comments. **(1)** To see what the arguments of the integral are, try to see in Figure 6.1 what affects its value. If we change any of a, b, f, the integral will change. The argument x does not affect the value of the integral, even though it is present in the notation of the integral (6.1). For example, the same integral can be written as $\int_a^b f(t)dt$ and it will not depend on t. The *variable of integration* t is included here to facilitate some manipulations with integrals.

(2) Instead of changing separately a or b we can think of the whole segment $[a,b]$ as changing, and then we can ask what properties the integral has as a function of a segment (when f is fixed). Its fundamental and geometrically obvious property is *additivity with respect to the domain of integration*: if $a < c < b$, then

$$\int_a^b f(x)dx = \int_a^c f(x)dx + \int_c^b f(x)dx$$

(see Figure 6.1).

(3) Since the area of a segment is equal to zero, for any a we have

$$\int_a^a f(x)dx = 0. \tag{6.2}$$

(4) We can also ask how the integral changes when the segment $[a,b]$ is fixed and the function f changes. Let f,g be two functions and let c,d be two numbers. The function $cf(x) + dg(x)$ is called a *linear combination* of f,g with coefficients c,d. The *linearity* property states that

$$\int_a^b [cf(x) + dg(x)]dx = c\int_a^b f(x)dx + d\int_a^b g(x)dx.$$

This property is quite similar to linearity of means and its applications are along the same lines but the proof requires a different level of math. This is one of those cases when you can use the *cow-and-goat principle*: here is a cow and there is a goat; they are quite similar and can be used, to a certain extent, for the same purposes. When it's time to tell you how the cow is different from the goat, I will.

(5) The *order property* of integrals is immediate from their interpretation as area: if the graph of the function f is above that of the function g, $f(x) \geq g(x)$ for all $x \in [a,b]$, then

$$\int_a^b f(x)dx \geq \int_a^b g(x)dx.$$

Unit 6.2   Density of a continuous random variable (density, interval formula in terms of densities)

Recall the short definition of a discrete random variable: values plus probabilities. Regarding the probabilities, we have established two properties: they are percentages and they satisfy the completeness axiom. We want an equivalent of these notions and properties in the continuous case.

Definition. Let X be a continuous random variable and let F_X be its distribution function. If there exists a function p_X such that

$$F_X(x) = \int_{-\infty}^x p_X(t)dt \text{ for all } -\infty < x < \infty, \tag{6.3}$$

then we say that the function p_X is the *density* of X.

Comments. **(1)** Please follow my practice of attaching subscripts to emphasize that F_X and p_X pertain to X.

(2) The definition appeals to existence of a function p_X satisfying (6.3). All continuous random variables that do not satisfy this existence requirement are left out in this course.

Property 1. From (6.3) we can obtain a more general equation (an *interval formula in terms of densities*):

$$P(a < X \leq b) = \int_a^b p_X(t)dt. \qquad (6.4)$$

Proof. Use the interval formula (5.16) and additivity of integrals to get

$$P(a < X \leq b) = F_X(b) - F_X(a) = \int_{-\infty}^b p_X(t)dt - \int_{-\infty}^a p_X(t)dt = \int_a^b p_X(t)dt.$$

Property 2. The probability of X taking any specific value is zero because by (6.2)

$$P(X = a) = P(X \leq a) - P(X < a) = \int_a^a p_X(t)dt = 0.$$

For this reason $P(X \leq x)$ is the same as $P(X < x)$ and $P(a < X \leq b)$ is the same as $P(a \leq X \leq b)$.

Property 3. $p_X(t) \geq 0$ for all t.

Proof. If the density is negative at some point x_0, we can take a and b close to that point and then by the interval formula (6.4) $P(a < X \leq b)$ will be negative, which is impossible.

Property 4. Any density satisfies the completeness axiom in the form

$$\int_{-\infty}^{+\infty} p_X(t)dt = 1$$

(the total area under the graph is 1).

Proof. This is because $\{-\infty < X < +\infty\}$ is a sure event and $F_X(+\infty) = 1$.

Differences between the cow and goat. **(i)** In the discrete case the variable takes its values with positive probabilities $P(X = x_i) = p_i$. In the continuous case $P(X = x) = 0$ for any x.

(ii) In the discrete case the numbers p_i are probabilities (percentages). In the continuous case the values of the density $p_X(t)$ mean nothing and can be larger than 1. It's the area $\int_a^b p_X(t)dt$ that means probability!

Exercise 6.1 Take two related continuous random variables and draw what you think would be their densities. For example, you can take: (i) distributions of income in a wealthy neighborhood and in a poor neighborhood, (ii) distributions of temperature in winter and summer in a given geographic location; (iii) distributions of electricity consumption in two different locations at the same time of the year. Don't forget that the total density is 1!

Unit 6.3 Mean value of a continuous random variable (mean)

This whole section is about the definition of EX. In the preamble of the definition you will see the argument of type used in the times of Newton and Leibniz, the fathers of calculus. It is still used today in inductive arguments by modern mathematicians.
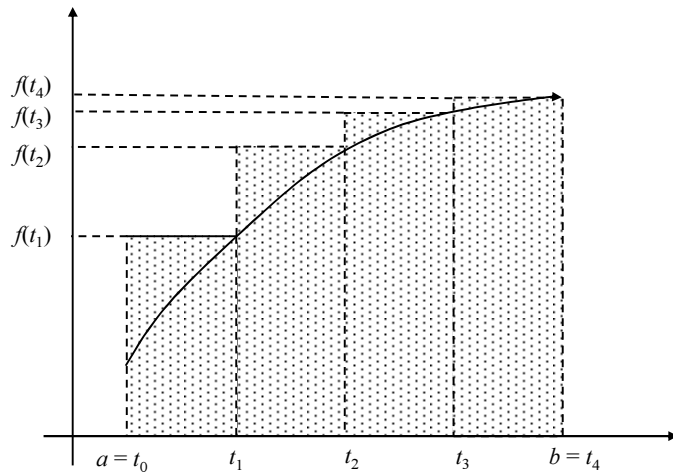
Figure 6.2 Approximation of an integral

Approximation of integrals by integral sums. The idea of the next construction is illustrated in Figure 6.2. The segment $[a,b]$ is divided into many small segments, denoted

$$[t_0,t_1], ..., [t_{m-1},t_m] \tag{6.5}$$

where the leftmost point t_0 coincides with a and the rightmost point t_m is b. By additivity of integrals, the integral over the whole segment is a sum of integrals over subsegments:

$$\int_a^b f(x)dx = \int_{t_0}^{t_1} f(x)dx + ... + \int_{t_{m-1}}^{t_m} f(x)dx. \tag{6.6}$$

When m is large and the subsegments are small, for each of them the area $\int_{t_{i-1}}^{t_i} f(x)dx$ of a curvilinear figure can be approximated by the area $f(t_i)(t_i - t_{i-1})$ of a rectangle. Therefore (6.6) is approximately

$$\int_a^b f(x)dx \approx f(t_1)(t_1 - t_0) + ... + f(t_m)(t_m - t_{m-1}). \tag{6.7}$$

In particular, if the segment $[a,b]$ itself is small, then

$$\int_a^b f(x)dx \approx f(b)(b-a). \tag{6.8}$$

Preamble. Let X be a continuous random variable that takes values in $[a,b]$. Let us divide $[a,b]$ into equal parts (6.5). If we have a sample of n observations on X, they can be grouped in batches depending on which subsegment they fall into. Let n_1 be the number of observations falling into the first subsegment $[t_0,t_1]$, ..., n_m the number of observations

falling into the last one $[t_{m-1}, t_m]$. All observations falling into $[t_{i-1}, t_i]$ are approximately equal to the right end t_i of this subsegment. Then $n_1 + ... + n_m = n$ and by the mean of grouped data formula (see the third line in (5.2))

$$EX \approx \frac{n_1 t_1 + ... + n_m t_m}{n}.$$

(6.9)

In the frequentist interpretation, $\dfrac{n_i}{n}$ approximates the probability that X takes values in $[t_{i-1}, t_i]$. By the interval formula (6.4) and approximation (6.8)

$$\frac{n_i}{n} \approx P(t_{i-1} < X \le t_i) = \int_{t_{i-1}}^{t_i} p_X(t) dt$$

$$\approx p_X(t_{i-1})(t_i - t_{i-1})$$

(6.10)

when the subsegments are short enough. (6.9) and (6.10) imply that

$$EX \approx t_1 p_X(t_1)(t_1 - t_0) + ... + t_m p_X(t_m)(t_m - t_{m-1}).$$

(6.11)

From (6.7) and (6.11) we see that

$$EX \approx \int_a^b t p_X(t) dt$$

where the approximation improves as the partitioning into subsegments becomes finer.

Definition. The *mean of a continuous random variable* is defined by

$$EX = \int_a^b t p_X(t) dt.$$

Extending the density by zero outside $[a, b]$ we can with equal success define the mean by

$$EX = \int_{-\infty}^{+\infty} t p_X(t) dt.$$

(6.12)

Unit 6.4   Properties of means, covariance, variance and correlation

It might be relieving to know that it's only the definition of the mean that is different from the discrete case. All properties of means, covariance and variance derived from that definition are absolutely the same as in the discrete case. Here is the list of properties that will be constantly used:

(1) *Linearity of means*: $E(aX + bY) = aEX + bEY$.

(2) *Expected value of a constant*: $Ec = c$.

(3) *Multiplicativity of expected values*: $EXY = (EX)(EY)$ for independent X, Y.

(4) *Linearity of covariance*: for example, for the first argument
$\operatorname{cov}(aX + bY, Z) = a\operatorname{cov}(X, Z) + b\operatorname{cov}(Y, Z).$

(5) *Alternative expression for covariance*: $\operatorname{cov}(X, Y) = EXY - (EX)(EY)$.

(6) *Independent variables are uncorrelated*: $\text{cov}(X,Y) = 0$.

(7) *Covariance with a constant*: $\text{cov}(X,c) = 0$.

(8) *Symmetry of covariance*: $\text{cov}(X,Y) = \text{cov}(Y,X)$.

(9) *Link between variance and covariance*: $V(X) = \text{cov}(X,X)$.

(10) *Alternative expression for variance*: $V(X) = EX^2 - (EX)^2$.

(11) *Variance of a linear combination*: $V(aX + bY) = a^2 V(X) + 2ab\,\text{cov}(X,Y) + b^2 V(Y)$.

(12) *Variance of a constant*: $V(c) = 0$.

(13) *Nonnegativity*: $V(X) \geq 0$.

(14) *Variables with vanishing variance*: $V(X) = 0$ if and only if $X = c$.

(15) *Additivity of variance*: $V(X + Y) = V(X) + V(Y)$ for independent X, Y.

(16) *Standard deviation is homogeneous*: $\sigma(aX) = |a|\,\sigma(X)$.

(17) *Range of the correlation coefficient*: $-1 \leq \rho(X,Y) \leq 1$ for any X, Y.

(18) *Standardization*: $Y = \dfrac{X - EX}{\sigma(X)}$ has mean zero and unit variance.

Here only property (3) requires additional definitions: joint density, independence etc. Also, for (10) we need the definition of a function of a random variable, and this is better explained following the cow-and-goat principle.

Table 6.1 Definition of a function of a random variable

	The variable has values	Probability/Density	A *function of a variable* has values
Discrete case	x_i	p_i	$f(x_i)$
Continuous case	t	$p_X(t)$	$f(t)$

Example 6.1 X^2 takes values t^2. $\sin X$ takes values $\sin t$. Therefore

$$EX^2 = \int_{-\infty}^{+\infty} t^2 p_X(t)dt, \quad E\sin X = \int_{-\infty}^{+\infty} (\sin t) p_X(t)dt.$$

Unit 6.5 The uniform distribution (uniformly distributed variable, primitive function, integration rule)

You might want to review the definition of a uniformly distributed discrete variable in Unit 5.2.

Example 6.2 Take the power cable of your computer and stretch it along the real line. How would you describe the probability of the cable failing at any fixed point (plugs don't count)? Do you think this probability at some points is higher than at others?

Intuitive definition. Fix some segment $[a,b]$. We say that a random variable U is *uniformly distributed* over $[a,b]$ if

        (a) U does not take values outside $[a,b]$ and

(b) its values inside $[a,b]$ are equally likely.

After having heard the motivating example and the intuitive definition most students are able to formally define the density of U.

Formal definition. U is such a variable that its density is

(a) null outside $[a,b]$ and

(b) constant inside $[a,b]$.

Denoting

$$p_U(t) = \begin{cases} 0 & \text{if } t \notin [a,b] \\ 1 & \text{if } t \in [a,b] \end{cases}$$

we can find the constant using the fact that the total density must be 1:

$$\int_{-\infty}^{+\infty} p_U(t)dt \qquad\qquad \text{(using additivity of integrals)}$$

$$= \int_{-\infty}^{a} p_U(t)dt + \int_{a}^{b} p_U(t)dt + \int_{b}^{+\infty} p_U(t)dt \quad \text{(using density definition)}$$

$$= \int_{a}^{b} cdt \qquad\qquad\qquad \text{(this is area of a rectangle)}$$

$$= c(b-a) = 1$$

so $c = \dfrac{1}{b-a}$. The result is

$$p_U(t) = \begin{cases} 0 & \text{if } t \notin [a,b] \\ \dfrac{1}{b-a} & \text{if } t \in [a,b] \end{cases} \qquad (6.13)$$

Important. In the interval formulas (5.16) and (6.4) the segment $[a,b]$ is arbitrary, whereas in the definition of U it is fixed. To remember this, sometimes it's useful to denote $U_{[a,b]}$ the variable which is uniformly distributed over $[a,b]$.

Property 1. Usually students can guess that $EU_{[a,b]} = \dfrac{a+b}{2}$.

Proof. Using definitions (6.12) and (6.13) we get

$$EU_{[a,b]} = \int_{-\infty}^{+\infty} tp_U(t)dt = \int_{a}^{b} tp_U(t)dt = \frac{1}{b-a}\int_{a}^{b} tdt. \qquad (6.14)$$

The last integral can be found using two steps from calculus.

Step 1. We have to find what is called a primitive function. $F(x)$ is called a *primitive function* of $f(x)$ if $F'(x) = f(x)$. The primitive of $f(t) = t$ is $F(t) = \dfrac{t^2}{2}$.

Step 2 is to use the primitive according to the *integration rule*

$$\int_a^b f(t)dt = F(b) - F(a).$$

In our case $\int_a^b t\,dt = \dfrac{b^2}{2} - \dfrac{a^2}{2} = \dfrac{(b-a)(a+b)}{2}$. This together with (6.14) implies

$$EU_{[a,b]} = \frac{1}{b-a}\frac{(b-a)(a+b)}{2} = \frac{a+b}{2}. \tag{6.15}$$

Property 2. $V\left(U_{[a,b]}\right) = \dfrac{(b-a)^2}{12}$.

Proof. There is no need to try to guess or remember this result. We start with the alternative expression:

$$V\left(U_{[a,b]}\right) = EU_{[a,b]}^2 - \left(EU_{[a,b]}\right)^2. \tag{6.16}$$

Here the first term is

$$EU_{[a,b]}^2 = \int_{-\infty}^{+\infty} t^2 p_U(t)dt = \frac{1}{b-a}\int_a^b t^2 dt.$$

The primitive of $f(t) = t^2$ is $F(t) = \dfrac{t^3}{3}$. We can continue the previous display as

$$EU_{[a,b]}^2 = \frac{1}{b-a}\left(\frac{b^3}{3} - \frac{a^3}{3}\right) = \frac{1}{b-a}\frac{(b-a)(b^2+ab+a^2)}{3} = \frac{b^2+ab+a^2}{3}. \tag{6.17}$$

The equation $b^3 - a^3 = (b-a)(b^2 + ab + a^2)$ we have used can be verified directly. Combining (6.15), (6.16) and (6.17) yields

$$V\left(U_{[a,b]}\right) = \frac{b^2+ab+a^2}{3} - \left(\frac{a+b}{2}\right)^2 = \frac{b^2-2ab+a^2}{12} = \frac{(b-a)^2}{12}.$$

Exercise 6.2 The distribution function of $U_{[a,b]}$ is

$$F_U(t) = \begin{cases} 0 & \text{if } t < a \\ \dfrac{t-a}{b-a} & \text{if } a \le t \le b \\ 1 & \text{if } t > b. \end{cases}$$

Unit 6.6 The normal distribution (parameter, standard normal distribution, normal variable and alternative definition, empirical rule)

6.6.1 Parametric distributions

Following the general recommendation to think in terms of sets rather than elements, let us think about families of similar distributions. One element of a family of similar distributions can be thought of as a function. The argument of the function is called a *parameter*. This

parameter can be a number or a vector. All families we have seen so far are either one-parametric or two-parametric. Let us look at the examples.

(i) The family of Bernoulli distributions is one-parametric. A Bernoulli distribution is uniquely identified by the probability $p \in (0,1)$.

(ii) The binomial variable has two parameters: n and p. It's value x which we write in the expression $P(S_n = x)$ is not a parameter.

(iii) The Poisson distribution has one parameter: $EX = V(X) = \lambda$.

(iv) The uniform distribution has two parameters, which are the endpoints of the segment $[a,b]$.

(v) Remembering about parameters is especially important when you work with more than one distribution. For example, the result from Unit 5.15 on Poisson approximation to the binomial distribution is clearer if you write it like this:

$$P(\text{Binomial} = x \,|\, n,\, p) \approx P(\text{Poisson} = x \,|\, \lambda = np).$$

6.6.2 Standard normal distribution

Definition. A *standard normal distribution*, denoted z, has a density function

$$p_z(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}. \tag{6.18}$$

It is universal in the sense that many other distributions converge to it. The exact meaning of this will be explained in **Theorem 6.1**.

We say that a function $f(x)$ of a real argument is *odd* if $f(-x) = -f(x)$. Such a function is called *symmetric* about zero or *even* if $f(-x) = f(x)$. We shall need a simple statement: If a function $f(x)$ is symmetric, then $xf(x)$ is odd: $(-x)f(-x) = -xf(x)$.

Property 1. The density of a standard normal is symmetric about zero.

Proof. $p_z(-t) = \dfrac{1}{\sqrt{2\pi}} e^{-\frac{(-t)^2}{2}} = p_z(t)$ for $t > 0$.

Property 2. $Ez = 0$.

Proof. The symmetry of the density implies that $tp_z(t)$ is an odd function and by additivity of integrals

$$Ez = \int_R tp_z(t)dt = \int_{-\infty}^0 tp_z(t)dt + \int_0^{+\infty} tp_z(t)dt.$$

This mean is zero because $A = \int_0^{+\infty} tp_z(t)dt$ cancels out with $-A = \int_{-\infty}^0 tp_z(t)dt$, see **Figure 6.3**.
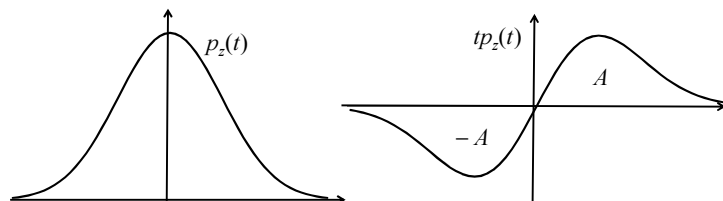
Figure 6.3 Density of standard normal

Property 3. $Ez^2 = 1$.

We take this for granted. But note that $V(z) = Ez^2 - (Ez)^2 = 1$ and $\sigma(z) = 1$.

Property 4. Even though the standard normal can take arbitrarily large values, the probability of z taking large values quickly declines. More precisely,

$$P(|z| > 1) = 0.3174, \ P(|z| > 2) = 0.0456, \ P(|z| > 3) = 0.0028. \tag{6.19}$$

Proof. Using the definition of density

$$P(|z| > 1) = \int_{-\infty}^{-1} p_z(t)dt + \int_{1}^{+\infty} p_z(t)dt \qquad \text{(by symmetry)}$$

$$= 2\int_{1}^{+\infty} p_z(t)dt = \frac{2}{\sqrt{2\pi}} \int_{1}^{\infty} e^{-\frac{t^2}{2}} dt.$$

There are no simple rules to calculate integrals like this. Therefore such integrals have been tabulated. Whatever density and table you deal with, try to express the integral through the area given in the table. Table 1 on p.841 in NCT gives the area $F(z)$ for $z > 0$ (in that table z means a real number, not the standard normal variable). By the complement rule and mentioned table

$$\int_{1}^{+\infty} p_z(t)dt = 1 - \int_{-\infty}^{1} p_z(t)dt = 1 - F(1) = 1 - 0.8413$$

so $P(|z| > 1) = 2(1 - 0.8413) = 0.3174$. In the other two cases the derivation is similar.

6.6.3 Normal distribution

Definition. A *normal variable* X is defined as a linear transformation of the standard normal: $X = \sigma z + \mu$, where $\sigma > 0$.

Property 1. The density of X is bell-shaped and centered on μ: $EX = \mu$.

Proof. By linearity of means $EX = \sigma Ez + \mu = \mu$. Recall our discussion of linear transformations: multiplication of z by σ stretches/squeezes its density and the subsequent addition of μ moves the result to the left or right.

Property 2. $V(X) = \sigma^2$.

Proof. Adding a constant doesn't change a variable and variance is homogeneous: $V(X) = V(\sigma z + \mu) = V(\sigma z) = \sigma^2 V(z) = \sigma^2$.

Property 3. Standardization of a normal distribution gives a standard normal distribution.

Proof. By properties 1 and 2 $\dfrac{X - \mu}{\sigma} = \dfrac{\sigma z + \mu - \mu}{\sigma} = z$.

Property 4. *Empirical rule*:

$$P(|X-\mu|\le\sigma)=0.68,\ P(|X-\mu|\le 2\sigma)=0.95,\ P(|X-\mu|\le 3\sigma)=0.997.$$

Proof. This follows from (6.19). For example, $P(|X-\mu|\le\sigma)=P(|z|\le 1)=1-0.3174\approx 0.68$.

Notation. Another notation for a normal variable with mean μ and variance σ^2 is $N(\mu,\sigma^2)$. The standard normal in this notation is $N(0,1)$. We see that the family of normal distributions is two-parametric. Each of the parameters takes an infinite number of values, and it would be impossible to tabulate all normal distributions. This is why we need the standardization procedure.

Definition from NCT. X is *normal* if its density is $p_X(t)=\dfrac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2}$. Working with this definition is more difficult, trust me.

Unit 6.7 Normal approximation to the binomial (point estimate, interval estimate, confidence interval, confidence level, left and right tails, significance level, convergence in distribution, central limit theorem)

Example 6.3 Compare two statements:

(i) The price of Microsoft stock tomorrow is expected to be \$43.

(ii) With probability 0.95 the price of Microsoft stock tomorrow will be between \$40 and \$46, $P(40\le X\le 46)=0.95$.

The first statement, called a *point estimate*, entails too much uncertainty. The second estimate, called an *interval estimate*, is preferable.

Definition. An interval which contains the values of a random variable X with high probability, $P(a\le X\le b)=1-\alpha$, where α is close to zero, is called a *confidence interval* and the number $1-\alpha$ is called a *confidence level*. The areas $\{X<a\}$ and $\{X>b\}$ are called a *left tail* and *right tail*, respectively. Confidence intervals can be two-tail and one-tail. If both probabilities $P(X<a)$ and $P(X>b)$ are positive, it is a *two-tail interval*; if one of them is zero, we deal with a *one-tail interval*. The number α is called a *significance level*.

The interest in confidence intervals leads us to the next definition.

Definition. Let $X_1,\ X_2,\ ...$ be a sequence of random variables. We say that it *converges* to a random variable X *in distribution* if $F_{X_n}(x)$ approaches $F(x)$ for all real x.

Corollary 6.1 If X_n converges to X in distribution, then for all large n by the interval formula

$$P(a\le X_n\le b)=F_{X_n}(b)-F_{X_n}(a)\approx F_X(b)-F_X(a)=P(a\le X\le b). \qquad (6.20)$$

Theorem 6.1 (*Central limit theorem*) Let $X_1,\ X_2,\ ...$ be i.i.d. random variables obtained by sampling from an arbitrary population. We want to obtain confidence intervals for the sums $S_n=X_1+...+X_n$. Denote

$$z_n=\frac{S_n-ES_n}{\sigma_{S_n}}$$

the standardized versions of S_n. z_n in general are not standard normal but they converge in distribution to a standard normal distribution.

Corollary 6.2 Let S_n be a binomial distribution. Then for any numbers a, b and large n one has

$$P(a \le S_n \le b) \approx P\left(\frac{a - ES_n}{\sigma_{S_n}} \le z \le \frac{b - ES_n}{\sigma_{S_n}} \right). \tag{6.21}$$

Proof. We need two properties of ordering of real numbers:

(1) the inequality $u \le v$ is equivalent to the inequality $u + c \le v + c$ for any constant c,

(2) the inequality $u \le v$ is equivalent to the inequality $uc \le vc$ for any positive constant c.

It follows that the next three inequalities are equivalent:

(i) $a \le S_n \le b$, (ii) $a - ES_n \le S_n - ES_n \le b - ES_n$, (iii) $\dfrac{a - ES_n}{\sigma_{S_n}} \le \dfrac{S_n - ES_n}{\sigma_{S_n}} \le \dfrac{b - ES_n}{\sigma_{S_n}}$.

Hence, probabilities of these events are the same. Now by the central limit theorem

$$P(a \le S_n \le b) = P\left(\frac{a - ES_n}{\sigma_{S_n}} \le z_n \le \frac{b - ES_n}{\sigma_{S_n}} \right) \approx P\left(\frac{a - ES_n}{\sigma_{S_n}} \le z \le \frac{b - ES_n}{\sigma_{S_n}} \right).$$

Exercise 6.3 Obtain an approximate confidence interval for $\hat{p} = \dfrac{S_n}{n}$.

Unit 6.8   Joint distributions and independence (joint distribution function, marginal distribution function, independent variables: continuous case)

All definitions in this section had precedents.

Consider vectors (X, Y) and (x, y) ((X, Y) is a pair of random variables and (x, y) is a pair of real numbers). The function $F_{(X,Y)}(x, y) = P(X \le x, Y \le y)$ is called a *joint distribution function* of the vector (X, Y). The *marginal distribution functions* are defined to be $F_{(X,Y)}(x, \infty)$, $F_{(X,Y)}(\infty, y)$. They satisfy $F_{(X,Y)}(x, \infty) = F_X(x)$, $F_{(X,Y)}(\infty, y) = F_Y(y)$ (marginal distributions coincide with own distributions). We say that a function $p_{(X,Y)}(t_1, t_2)$ is a joint density of (X, Y) if

$$F_{(X,Y)}(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} p_{(X,Y)}(t_1, t_2) dt_1 dt_2.$$

Suppose X and Y have densities p_X and p_Y, respectively. X and Y are called *independent* if the joint density decomposes into a product of own densities: $p_{(X,Y)}(t_1, t_2) = p_X(t_1) p_Y(t_2)$. With these definitions we can prove Property 3) from Unit 6.4.

Exercise 6.4 For independent continuous random variables we have multiplicativity of means: $EXY = (EX)(EY)$.

Proof. Everything follows from definitions:

$$EXY = \int_R \int_R t_1 t_2 p_{(X,Y)}(t_1, t_2) dt_1 dt_2 \qquad \text{(by independence)}$$

$$= \int_R \int_R t_1 t_2 p_X(t_1) p_Y(t_2) dt_1 dt_2$$

$$= \int_R t_1 p_X(t_1) dt_1 \int_R t_2 p_Y(t_2) dt_2 = (EX)(EY).$$

Constructing independent standard normal variables. We wish to have independent standard normal z_1, z_2 (both must have the same density $p_z(t)$). To satisfy the definition of independence just put $p_{(z_1, z_2)}(t_1, t_2) = p_{z_1}(t_1) p_{z_2}(t_2)$. In a similar way for any natural n we can construct n independent standard normal variables.

Unit 6.9   Questions for repetition

1. Discuss the five properties of integrals that follow their definition: do summations have similar properties?

2. How do you prove that a density is everywhere nonnegative?

3. How do you justify the definition of the mean of a continuous random variable?

4. With the knowledge you have, which of the 18 properties listed in Unit 6.4 can you prove? Mark with a star those you can't prove at this point and look for the answers later.

5. Define the uniformly distributed random variable. Find its mean, variance and distribution function.

6. In one block give the properties of the standard normal distribution, with proofs, where possible, and derive from them the properties of normal variables.

7. What is the relationship between distribution functions F_z, F_X if $X = \sigma z + \mu$? How do you interpret $F_z(1) - F_z(-1)$?

8. Divide the statement of the central limit theorem into a preamble and convergence statement. Provide as much justification/intuition for the preamble as you can. How is this theorem applied to the binomial variable?

9. Draw a parallel between the proofs of multiplicativity of means for discrete and continuous random variables.

10. Minimum required: Exercises 6.5, 6.15, 6.25, 6.45, 6.85, 6.95.

Chapter 7   Sampling and Sampling Distribution

Unit 7.1   Sampling distribution (simple random sample, quota sampling, statistic, sampling distribution, estimator, unbiasedness, upward and downward bias,)

Example 7.1 Suppose we are interested in finding the average income of the population of Kazakhstan. The random variable is income of a person (incomes have frequencies). Its mean is the parameter of the population we are interested in. Observing the whole population is costly. Therefore we want to obtain a sample and make inference (conclusions) about the population from the sample information. Review the theory we've studied so far and satisfy yourself that most theoretical facts critically depend on two assumptions: randomness and independence of observations.

Definition. A *simple random sample* satisfies two conditions:

(i) Every object has an equal probability of being selected and

(ii) The objects are selected independently.

Comments. **(1)** Condition (i) can be satisfied by shuffling in case of playing cards. If a list of objects is available, it can be satisfied using random numbers generated on the computer. Suppose we have a list of N objects to choose from. The function `RAND()` (with empty parentheses) from Excel is uniformly distributed over the segment (0,1). Then `N*RAND()` is uniformly distributed over $(0, N)$. Applying the function `INT` which rounds a number down to the closest integer we get `INT(N*RAND())` that takes integer values from 0 to $N-1$.

(2) It's good to know a couple of examples when the above conditions are violated. One example is sampling without replacement (see section 5.11.1). Another is *quota sampling*. In quota sampling elements are sampled until certain quotas are filled. For example, if the percentage of men in the country is 49 and the percentage of employed is 80, one might want to keep the same percentages in the sample. The problem with this type of sampling is dependence (inclusion of a new element in the sample depends on the previously selected ones) and impossibility to satisfy more than one quota simultaneously, unless by coincidence.

Definition. Let τ be a population parameter (for example, $\tau = \mu_X$ can be the population mean) and let T be the sample characteristic designed to estimate τ (for example, the sample mean is a good estimator for the population mean because $E\overline{X} = \mu_X$). T is constructed from sample observations $X_1, ..., X_n$ and in theory these observations are thought of as ex-ante random variables. Therefore T is a random variable and, as any random variable, it has a distribution. In the situation just described T is called a *statistic* and its distribution is called a *sampling distribution*.

Working definition. A *statistic* is a random variable constructed from sample observations and designed to estimate a parameter of interest. A *sampling distribution* is a table "values + probabilities" for the statistic.

Definition. Let T be a statistic designed to estimate (to approximate) a population parameter τ. T is called an *estimator* of τ. To recapitulate, it is a random variable and it depends on the sample data. We say that T is an *unbiased estimator* of τ if $ET = \tau$. If $ET > \tau$ we say that T is *biased upward* and if $ET < \tau$ it is *biased downward*.

Comments. **(1)** If we think of τ as the target, then the equation $ET = \tau$ means that T hits the target on average (in a large number of trials).

(2) From (5.13) we know that the sample mean is an unbiased estimator of the population mean. In general, it is believed that for any population parameter there exists a statistic that estimates it. For example, the sample variance s_X^2 is an unbiased estimator of the population variance σ_X^2:

$$Es_X^2 = \sigma_X^2. \tag{7.1}$$

This is the true reason of definition (3.4).

Controlling variances in sampling is important for at least three reasons. Firstly, in a production process of a certain product large deviations of its characteristics from technological requirements are undesirable. Secondly, the statistical procedure for comparing sample means from two populations requires knowledge of corresponding variances. Thirdly, in survey sampling by reducing the sample variance we reduce the cost of collecting the sample.

Exercise 7.1 Study the proof of (7.1) on pp. 273-274 of NCT and work out a version of the proof for yourself with two purposes in mind: (i) you should be able to reproduce the proof with complete explanations and (ii) try to extract general ideas that can be applied in other situations.

Solution. Your solution may be very different from mine. From the proof we see that the variables must be i.i.d.

<u>Step 1</u>. The beginning of the proof reminds me the derivation of the alternative expression for variance. Let us write the alternative expression formula as $E(Y - EY)^2 = EY^2 - (EY)^2$ and let us apply it to a random variable Y with values $Y_1, ..., Y_n$ and uniform distribution $p_i = 1/n$. Then from the above formula we get $\dfrac{1}{n}\sum (Y_i - \bar{Y})^2 = \dfrac{1}{n}\sum Y_i^2 - \bar{Y}^2$. Putting here $Y_i = X_i - \mu_X$ we get $\bar{Y} = \bar{X} - \mu_X$ and

$$\frac{1}{n}\sum (X_i - \bar{X})^2 = \frac{1}{n}\sum (X_i - \mu_X)^2 - (\bar{X} - \mu_X)^2. \tag{7.2}$$

This is an equivalent of the first step in NCT. This trick will prove useful again in Chapter 12. Introducing μ_X here is necessary because in $V(X) = E(X - \mu_X)^2$ we need deviations from the population mean.

Always try to see general features in particular facts.

<u>Step 2</u>. We apply the population mean to both sides of (7.2). It's important to distinguish the population mean from the sample mean. Since X_i are identically distributed we have

$$E(X_i - \mu_X)^2 = E(X - \mu_X)^2 = \sigma_X^2. \tag{7.3}$$

Now equations (7.2), (7.3), (5.13) and the definition of sample variance imply

$$Es_X^2 = E\frac{1}{n-1}\sum(X_i - \bar{X})^2 = \frac{n}{n-1}E\frac{1}{n}\sum(X_i - \bar{X})^2$$

$$= \frac{n}{n-1}\left[\sigma_X^2 - \frac{1}{n}\sigma_X^2\right] = \frac{n}{n-1}\frac{n-1}{n}\sigma_X^2 = \sigma_X^2.$$

Exercise 7.2 Prove the sample covariance $s_{X,Y}$ is an unbiased estimator of the population covariance $cov(X,Y)$.

Unit 7.2 Sampling distributions of sample variances (chi-square distribution, degrees of freedom)

Definition. Let $z_1,...,z_n$ be standard normal independent (we know from Unit 6.8 how to construct them). Then the variable $\chi_n^2 = \sum_{i=1}^{n} z_i^2$ is called a *chi-square distribution* with n *degrees of freedom*. "Degrees of freedom" is just a formal parameter that determines which entry to look up in a statistical table.

Property 1. χ_n^2 is nonnegative, therefore its density $p_{\chi_n^2}(t)$ vanishes for $t < 0$.

Property 2. $E\chi_n^2 = n$.

Proof. $E\chi_n^2 = \sum_{i=1}^{n} Ez_i^2 = n$ by linearity of means and Property 3 of standard normals (section 6.6.2).

Property 3. $V(\chi_n^2) = 2n$.

Proof. By independence and the alternative expression for variance

$$V(\chi_n^2) = \sum_{i=1}^{n} V(z_i^2) = \sum_{i=1}^{n} [Ez_i^4 - (Ez_i^2)^2].$$

We accept without proof that

$$Ez^4 = 3 \text{ for a standard normal.}$$

Hence, $V(\chi_n^2) = \sum_{i=1}^{n} (3-1) = 2n$.

Usually I use the name "Theorem" for complex statements that will not be proved here.

Theorem 7.1 For a random sample of n independent observations from a normal population with variance σ_X^2 the variable

$$\chi_{n-1}^2 = \frac{(n-1)s_X^2}{\sigma_X^2}$$

is distributed as chi-square with $n-1$ degrees of freedom.

Exercise 7.3 Find $V(s_X^2)$.

Solution. By the preceding theorem, homogeneity of variance and Property 3

$$V(s_X^2) = V\left(\frac{\sigma_X^2}{n-1}\chi_{n-1}^2\right) = \frac{\sigma_X^4}{(n-1)^2}V\left(\chi_{n-1}^2\right) = \frac{2\sigma_X^4}{n-1}.$$

> As you get used to algebra, don't write every little detail. Try to perform more operations mentally. Tigran Petrosian became a chess grandmaster because at elementary school his teacher wanted him to write detailed solutions, while he wanted to give just an answer.

Exercise 7.4 Find a confidence interval for s_X^2.

Solution. Since s_X^2 is positive, in the expression $P(a \le s_X^2 \le b)$ it makes sense to consider only nonnegative a. You might want to review the proof of **Corollary 6.2**. By **Theorem 7.1**

$$P(a \le s_X^2 \le b) = P\left(\frac{(n-1)a}{\sigma_X^2} \le \frac{(n-1)s_X^2}{\sigma_X^2} \le \frac{(n-1)b}{\sigma_X^2}\right)$$

$$= P\left(\frac{(n-1)a}{\sigma_X^2} \le \chi_{n-1}^2 \le \frac{(n-1)b}{\sigma_X^2}\right).$$

If we want this probability to be $1-\alpha$, we can use Table 7 from NCT to find a, b so that both tails will have probability $\alpha/2$:

$$P\left(\chi_{n-1}^2 < \frac{(n-1)a}{\sigma_X^2}\right) = P\left(\chi_{n-1}^2 > \frac{(n-1)b}{\sigma_X^2}\right) = \frac{\alpha}{2}.$$

Unit 7.3   Monte Carlo simulations

Exercise 7.5 (Modeling Bernoulli) Let $U = U_{(0,1)}$ denote the variable that is uniformly distributed over (0,1). Put

$$X = \begin{cases} 0 & \text{if } U > p \\ 1 & \text{if } U \le p \end{cases} \tag{7.4}$$

Prove that then X is a Bernoulli variable and $P(X = 1) = p$.

Proof. X takes only values 0 and 1. By the definition of U we have

$P(X = 1) = P(U \le p) = \int_0^p dt = p.$

In Excel the function `RAND()` supplies the variable U. The mathematical definition (7.4) in Excel is realized by the command `IF(condition, `v_1`, `v_2`)`. If `condition` is true, it gives v_1, and if it is wrong, the command gives v_2. Thus, Exercise 2.1 models a Bernoulli variable with $p = 1/4$.

Exercise 7.6 (Modeling binomial) How would you model the binomial variable S_3 in Excel?

Solution. With three independent variables X_1, X_2, X_3 defined by (7.4) put $S_3 = X_1 + X_2 + X_3$. When you put the function `RAND()` in different cells, the resulting variables will be independent. Therefore Exercise 2.2 models the binomial for which $ES_3 = 3p = 3/4$.

Before we move on to modeling continuous random variables we need to review the notion of an inverse function. Let f be a function with the domain $D(f)$ and range $R(f)$. If for any value $y \in R(f)$ the equation $f(x) = y$ has a unique solution, we say that f has an *inverse function* f^{-1} and write the solution as $x = f^{-1}(y)$.

> In equations try to write the knowns on the right and the unknowns (or what you define) on the left side.

Working definition. f is *invertible* if the equality of values $f(x_1) = f(x_2)$ implies the equality of arguments $x_1 = x_2$.

Example 7.2 Consider the function $y = x^2$ on the real line. It does not have an inverse function because for any value $y > 0$ there are two different solutions of the equation $x^2 = y$. The same function on a smaller domain $D(f) = [0, \infty)$ has an inverse because a positive square root of a positive number is unique. Thus, a function may not be invertible globally (on the whole domain) but can be invertible locally (on a subset of a domain).

Property 1. The direct function maps the argument to the value: $y = f(x)$. The inverse function maps the value to the argument: $x = f^{-1}(y)$. A successive application of the two mappings amounts to an identical mapping: $f(f^{-1}(y)) = y$.

Property 2. f is called *nondecreasing* if $x_1 \leq x_2$ implies $f(x_1) \leq f(x_2)$. Suppose f is invertible. Then f is nondecreasing if and only if its inverse is nondecreasing.

Proof. Let f be nondecreasing and let us prove that its inverse is nondecreasing. Suppose the opposite: there exist $y_1 > y_2$ such that $f^{-1}(y_1) \leq f^{-1}(y_2)$. Then by monotonicity of the direct function $y_1 = f(f^{-1}(y_1)) \leq f(f^{-1}(y_2)) = y_2$ which contradicts our assumption. Hence, the assumption is wrong and the inverse is nondecreasing.

Conversely, let f^{-1} be nondecreasing and let us prove that f is nondecreasing. This follows from what we've proved and the fact that $(f^{-1})^{-1} = f$.

Condition sufficient for invertibility of a distribution function. The distribution function F_X of a random variable X is invertible where the density p_X is positive.

Proof. Suppose p_X is positive in some interval $[a, b]$ and let us show that the distribution function is invertible in that interval. Suppose $a \leq x_1 < x_2 \leq b$. Then by the interval formula $0 < \int_{x_1}^{x_2} p_X(t) dt = F_X(x_2) - F_X(x_1)$. Thus, the values $F_X(x_2), F_X(x_1)$ cannot be equal and by the working definition F_X^{-1} exists.

Corollary 7.1 For a normal variable its distribution function is invertible everywhere. The distribution function of the chi-square is invertible on the half-axis $[0, \infty)$. For $U_{(a,b)}$ its distribution function is invertible on $[a, b]$.

Exercise 7.7 How would you model an arbitrary random variable with an invertible distribution function?

Solution. Put $Y = F_X^{-1}(U)$ where U is uniformly distributed on (0,1). Since F_X is nondecreasing, the inequality $F_X^{-1}(U \le x)$ is equivalent to $U = F_X(F_X^{-1}(U)) \le F_X(x)$. Therefore for any real x $F_Y(x) = P(F_X^{-1}(U) \le x) = P(U \le F_X(x)) = \int_0^{F_X(x)} dt = F_X(x)$. Y is identically distributed with X.

Despite being short, this proof combines several important ideas. Make sure you understand them.

In Excel the function NORMINV(y , μ , σ) denotes $F_X^{-1}(y)$ where $X = \sigma z + \mu$ is normal. Hence, in Exercise 2.4 we model $X = 8z + 25$.

Unit 7.4   Questions for repetition

1. Fill out the next table:

Table 7.1 Comparison of population and sample formulas

	Population formula		Sample formula
Mean	Discrete and continuous		
Variance			
Standard deviation			
Correlation			

2. How would you satisfy the definition of the simple random sampling in case of measuring average income of city residents?

3. Define a sampling distribution and illustrate it with a sampling distribution of a sample proportion.

4. Stating all the necessary assumptions, derive the formulas for the mean and variance of the sample mean.

5. Suppose we sample from a Bernoulli population. How do we figure out the approximate value of p? How can we be sure that the approximation is good?

6. Define unbiasedness and prove that the sample variance is an unbiased estimator of the population variance.

7. Find the mean and variance of the chi-square distribution.

8. Find $V(s_X^2)$ and a confidence interval for s_X^2.

9. Minimum required: 7.9, 7.10, 7.30, 7.40, 7.57, 7.61.

Chapter 8 Estimation: Single Population

Unit 8.1 Unbiasedness and efficiency

In NCT this chapter starts with the definition of unbiasedness. In my exposition it was more appropriate to put that definition in Unit 7.1.

Exercise 8.1 If there exists one unbiased estimator of a population parameter, then there exist many unbiased estimators of that parameter.

Proof. Let $ET = \tau$. Take any random variable Y with mean zero and put $X = T + Y$. Then $EX = ET + EY = \tau$.

This exercise makes clear that even though unbiasedness is a desirable property, it is not enough to choose from many competing estimators of the same parameter. In statistics we want not only to be able to hit the target on average but also to reduce the spread of estimates around their mean.

Definition. If T_1 and T_2 are two unbiased estimators of the same parameter τ and $V(T_1) < V(T_2)$, then we say that T_1 is *more efficient*.

Exercise 8.2 If T_1 and T_2 are two unbiased estimators of the same parameter τ and $V(T_1) \neq V(T_2)$, then, with exception of two cases, there exists an unbiased estimator of the same parameter which is more efficient than either T_1 or T_2.

Proof. Define $T(a) = aT_1 + (1-a)T_2$ where a is real. This is a special type of a linear combination which is convenient to parameterize the points of the segment $[T_1, T_2]$. With $a = 0$ or 1 we get the left and right endpoints of the segment: $T(1) = T_1$, $T(0) = T_2$. Values $a \in (0,1)$ give all other points of the segment. Values $a \notin [0,1]$ give points of the straight line drawn through T_1, T_2.

$T(a)$ is unbiased: $ET(a) = aET_1 + (1-a)ET_2 = \tau$. Its variance is

$$V(T) = a^2 V(T_1) + 2a(1-a)\operatorname{cov}(T_1, T_2) + (1-a)^2 V(T_2)$$

and the first-order condition for its minimization is

$$2aV(T_1) + (2-4a)\operatorname{cov}(T_1, T_2) - 2(1-a)V(T_2) = 0.$$

Solving this equation for a,

$$a = \frac{V(T_2) - \operatorname{cov}(T_1, T_2)}{V(T_1)}.$$

The minimum of $V(T)$ is at point $a = 0$ if $V(T_2) = \operatorname{cov}(T_1, T_2)$ and at point $a = 1$ if $V(T_1) = \operatorname{cov}(T_1, T_2)$. In these two cases it is not possible to reduce variance by forming linear combinations. In all other cases $T(a)$ is more efficient than either T_1 or T_2.

Unit 8.2   Consistency (consistent estimator, Chebyshov inequality)

This is one of the topics most students don't understand, partly because too little information about it is given in NCT and partly because the notion is complex. I give more information but am not sure that it will simplify the topic. Consistency is an asymptotic property and therefore we have to work with a sequence of estimators.

Definition. Let $\{T_n\}$ be a sequence of estimators of the same parameter τ. We say that T_n is a *consistent estimator* of τ if the probability of deviations of T_n from τ decreases as n increases. More precisely, this condition is formulated as follows:

$$\text{For any } \varepsilon > 0, \ P(|T_n - \tau| \geq \varepsilon) \to 0 \text{ as } n \to \infty.$$

Comment. Let p_n denote the density of T_n. Then

$$P(|T_n - \tau| \geq \varepsilon) = \int_{-\infty}^{\tau-\varepsilon} p_n(t)dt + \int_{\tau+\varepsilon}^{+\infty} p_n(t)dt.$$
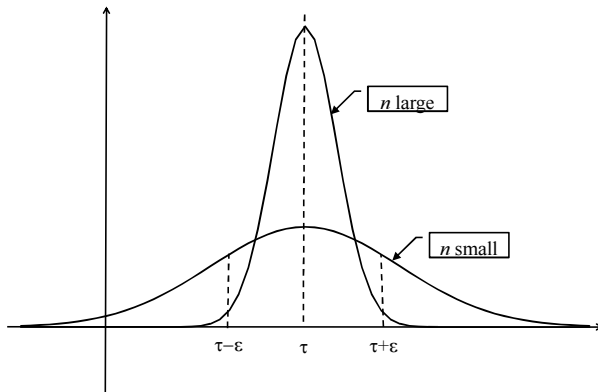
Figure 8.1 Illustration of consistency

In Figure 8.1 the area below the density outside the interval $(\tau-\varepsilon, \tau+\varepsilon)$ must tend to zero as $n \to \infty$, and this should be true for any $\varepsilon > 0$. The density of T_n collapses to a spike at τ.

Theorem 8.1 (*Chebyshov inequality*) For any continuous random variable X and for any $\varepsilon > 0$ one has $P(|X| \geq \varepsilon) \leq \dfrac{1}{\varepsilon^2} EX^2$.

Proof. In the area $|t| \geq \varepsilon$ the inequality $1 \leq \dfrac{t^2}{\varepsilon^2}$ is true. Therefore

$$P(|X| \geq \varepsilon) = \int_{-\infty}^{-\varepsilon} p_X(t)dt + \int_{\varepsilon}^{+\infty} p_X(t)dt \qquad \text{(using the above inequality)}$$

$$\leq \frac{1}{\varepsilon^2}\left(\int_{-\infty}^{-\varepsilon} t^2 p_X(t)dt + \int_{\varepsilon}^{+\infty} t^2 p_X(t)dt \right) \quad \text{(increasing the domain of integration)}$$

$$\leq \frac{1}{\varepsilon^2} \int_R t^2 p_X(t)dt = \frac{1}{\varepsilon^2} EX^2.$$

Consistency of the sample mean. If $X_1, X_2, \ldots$ are i.i.d., then $\bar{X}$ is a consistent estimator of μ_X.

Proof. By the Chebyshov inequality for any $\varepsilon > 0$

$$P(|\bar{X} - \mu_X| \geq \varepsilon) \leq \frac{1}{\varepsilon^2} E(\bar{X} - \mu_X)^2 = \frac{1}{\varepsilon^2} E(\bar{X} - E\bar{X})^2 = \frac{\sigma_X^2}{n\varepsilon^2} \to 0, \ n \to \infty.$$

Relationship between consistency and unbiasedness

Consistency is an asymptotic property and unbiasedness is a fixed-sample property. If we want to compare them, in both cases we have to assume an infinite sequence of estimators. Suppose we have a sequence of estimators $\{T_n\}$ of τ. We can ask what is the relationship between

(i) consistency of T_n and

(ii) unbiasedness of T_n for every n.

Exercise 8.3 We prove that (i) does not imply (ii). Put $T_n = \bar{X} + \frac{1}{n}$. The sample mean consistently estimates the population mean and $\frac{1}{n} \to 0$, therefore T_n consistently estimates the population mean but it is biased.

Exercise 8.4 We prove that (ii) does not imply (i). Suppose $\mu_X = 0$ and put $T_n = \bar{X}/(\sigma_X/\sqrt{n})$. We have $ET_n = 0$ (unbiasedness). By the central limit theorem T_n converges in distribution to standard normal z and for any $\varepsilon > 0$ $P(|z_n| \geq \varepsilon) \approx P(|z| \geq \varepsilon)$ which does not tend to zero.

An unbiased estimator sequence may have a large variance. When choosing between such a sequence and a consistent sequence we face a tradeoff: we can allow for a small bias if variance reduces significantly.

Unit 8.3 Confidence interval for the population mean. Case of a known σ (critical value, margin of error)

Assumptions. A sample of size n is drawn from a normal population. σ is assumed to be known, μ is unknown and the objective is to obtain a confidence interval for μ.

Step 1. Select the significance level α.

Step 2. Using Table 1 in the Appendix to NCT find a number $z_{\alpha/2} > 0$ such that

$$P(-z_{\alpha/2} < z < z_{\alpha/2}) = 1 - \alpha. \quad (8.1)$$

$z_{\alpha/2}$ is called a *critical value* of the z statistic corresponding to α (for a two-tail confidence interval). $\alpha/2$ is the probability of the left and right tails and the critical value could be equivalently defined by $\alpha/2 = P(z < -z_{\alpha/2}) = P(z > z_{\alpha/2})$.

Step 3. Let us standardize the sample mean

$$z = \frac{\overline{X} - \mu_{\overline{X}}}{\sigma_{\overline{X}}} = \frac{\overline{X} - \mu_X}{\sigma_X / \sqrt{n}} \qquad (8.2)$$

The second equality here uses (5.13). We want to prove that this is really a standard normal.

Theorem 8.2 A sum of independent normal variables is normal.

Exercise 8.5 A linear transformation of a normal variable is normal.

Proof. Suppose X is normal, $X = \sigma z + \mu$ where z is standard normal, and Y is its linear transformation, $Y = aX + b$. Then $Y = a(\sigma z + \mu) + b = a\sigma z + (a\mu + b)$ is normal as a linear transformation of a standard normal. Here the product $a\sigma$ may be negative but it doesn't matter because of the symmetry of z.

From **Theorem 8.2** it follows that $S_n = X_1 + ... + X_n$ is normal. (8.2) is a linear transformation of S_n, so it is normal too. And we know that, being a standardized variable, it has mean zero and variance 1. Thus, (8.2) is standard normal.

Now we plug (8.2) in (8.1) and use equivalent transformations to get

$$1 - \alpha = P\left(-z_{\alpha/2} < \frac{\overline{X} - \mu_X}{\sigma_X / \sqrt{n}} < z_{\alpha/2}\right) = P\left(-z_{\alpha/2}\frac{\sigma_X}{\sqrt{n}} < \overline{X} - \mu_X < z_{\alpha/2}\frac{\sigma_X}{\sqrt{n}}\right)$$

$$= P\left(\overline{X} - z_{\alpha/2}\frac{\sigma_X}{\sqrt{n}} < \mu_X < \overline{X} + z_{\alpha/2}\frac{\sigma_X}{\sqrt{n}}\right).$$

The quantity $ME = z_{\alpha/2}\dfrac{\sigma_X}{\sqrt{n}}$ is called a *margin of error*. We have obtained a confidence interval for the population mean

$$1 - \alpha = P\left(\overline{X} - ME < \mu_X < \overline{X} + ME\right). \qquad (8.3)$$

Note that (8.3) can be written in a different way:

$$1 - \alpha = P\left(\mu_X - ME < \overline{X} < \mu_X + ME\right). \qquad (8.4)$$

These two forms reflect different points of view:

(i) In (8.3) the population mean is unknown and estimated from the sample data.

(ii) In (8.4) the population mean is known and the purpose is to obtain ex-ante bounds for the sample mean.

Unit 8.4   Confidence interval for the population mean. Case of an unknown σ (t-distribution)

Assumptions. A sample of size n is drawn from a normal population. σ is unknown and estimated by the sample standard deviation, μ is unknown and the objective is to obtain a confidence interval for μ.

We apply a general statistical idea: when some parameter is unknown, replace it by its estimator. Applied to (8.2) this idea gives

$$t = \frac{\overline{X} - \mu_X}{s_X / \sqrt{n}}. \qquad (8.5)$$

This takes us to a new class of distributions.

Definition. Let $z_0, z_1, ..., z_n$ be standard normal independent. Then the random variable

$$t_n = \frac{z_0}{\sqrt{\sum_{i=1}^{n} z_i^2}}$$

is called a t-*distribution* with n degrees of freedom.

Property 1. Like z, t_n is symmetric around zero and therefore $Et_n = 0$.

Property 2. The tails of t_n are fatter than those of z. Both have densities that are bell-shaped, see Figure 8.7 in NCT. If the density of t_n is higher than that of z for large arguments, then for small arguments (around the origin) the opposite should be true because the total density is 1.

Property 3. As $n \to \infty$, t_n converges to z. This means that for values $n > 60$ that are not in Table 8 in the Appendix of NCT one can use values for z from Table 1.

Theorem 8.3 Under independent sampling from a normal population the variables $\bar{X}$ and s_X^2 are independent.

Exercise 8.6 Show that (8.5) is distributed as t_{n-1}.

Solution. Using **Theorem 7.1** we get

$$\frac{\bar{X} - \mu_X}{s_X / \sqrt{n}} = \frac{\bar{X} - \mu_X}{\sigma_X / \sqrt{n}} \frac{1}{\sqrt{s_X^2 / \sigma_X^2}} = \frac{z_0}{\sqrt{\chi_{n-1}^2 / (n-1)}}.$$

We have proved in Unit 8.3 that $z_0 = \dfrac{\bar{X} - \mu_X}{\sigma_X / \sqrt{n}}$ is standard normal. By **Theorem 8.3** and

definition of the t distribution (8.5) is really t_{n-1}.

Exercise 8.7 (1) Derive the confidence interval for μ_X when σ_X is unknown.

(2) What happens to the critical value $t_{n-1,\alpha/2}$ and to the confidence interval when (i) n increases, (ii) α decreases, (iii) s_X increases?

Unit 8.5   Confidence interval for population proportion

From (5.14) we know that $E\hat{p} = p$, $V(\hat{p}) = \dfrac{p(1-p)}{n}$. By the central limit theorem the variable

$$\frac{\hat{p} - E\hat{p}}{\sigma_{\hat{p}}} = \frac{\hat{p} - p}{\sqrt{\dfrac{p(1-p)}{n}}} \tag{8.6}$$

is approximately normal if n is large enough ($np(1-p) > 9$). The problem with (8.6) is that in the denominator we have an unknown p. Following the general statistical idea, we replace it by $\hat{p}$ and obtain

$$z_n = \frac{\hat{p} - p}{\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}}. \qquad (8.7)$$

Exercise 8.8 Using (8.7) derive the confidence interval for p. What is the expression for ME?

Unit 8.6   Questions for repetition

1. Show that existence of one unbiased estimator of some parameter implies existence of many unbiased estimators of the same parameter.

2. If you have two unbiased estimators of the same parameter, how do you construct a more efficient one?

3. Define consistency and illustrate geometrically.

4. Derive the confidence interval for the population mean when: (a) σ is unknown and (b) σ is unknown.

5. Prove that (8.5) is distributed as t_{n-1} stating all theoretical facts that you use.

6. Derive the confidence interval for population proportion. Why do you use z statistic in this case?

7. Minimum required: Exercises 8.13, 8.14, 8.25, 8.26, 8.33, 8.34.

Chapter 9   Estimation: Additional Topics

As we move into a more applied area, problem statements, their solutions and conditions for validity of the solutions will become more complex. The previous material should provide many enough stepping stones.

```
          ┌──────────────────┐
          │ Objective: compare│
          │ means of two samples│
          └──────────────────┘
```

Objective: compare means of two samples

Sample sizes are equal and the corresponding elements are similar (matched pairs)

Samples are very different (independent)

Population variances unknown

Case I. Population variances known

Case II. Population variances equal

Case III. Population variances unequal
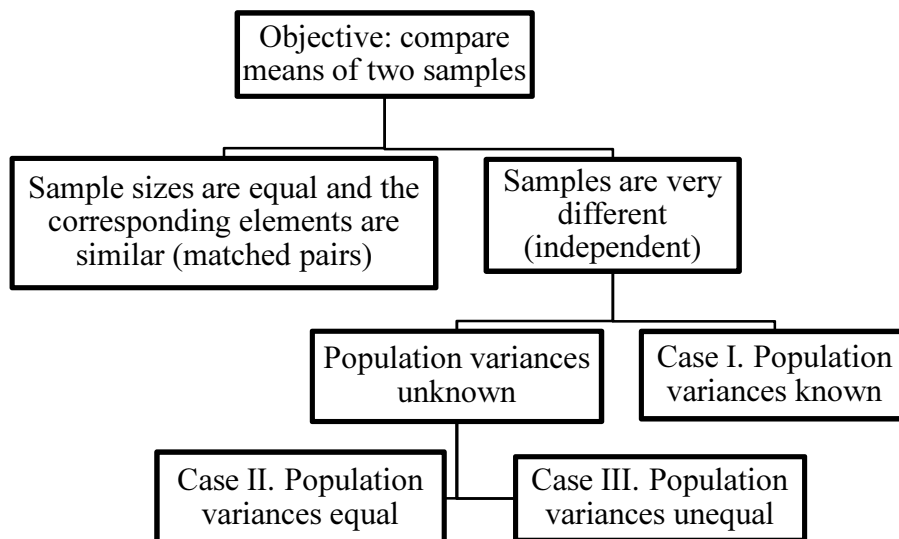
Unit 9.1   Matched pairs

Example 9.1 A group of students takes a course and their performance in the subject is measured before and after taking it. The sample sizes are equal and there is a high dependence between the grade x_i of the ith student before taking the course and the grade y_i after taking it. Instead of writing the samples as separate vectors $x = (x_1,...,x_n)$, $y = (y_1,...,y_n)$ we write the observations in pairs $(x_1, y_1),...,(x_n, y_n)$.

Assumptions. Suppose n matched pairs $(x_1, y_1),...,(x_n, y_n)$ of observations from populations with means μ_X, μ_Y are given and suppose that the differences $d_i = x_i - y_i$ are i.i.d. normal.

Finding the confidence interval

<u>Step 1</u>. Find the sample mean $\bar{d}$ and standard deviation s_d. Note that $\mu_d = \mu_X - \mu_Y$.

<u>Step 2</u>. Choose the level of significance α. Since we use sample standard deviation, the right statistic is t. Find its critical value for the two-tail confidence interval from the condition $P(t_{n-1} > t_{n-1,\alpha/2}) = \alpha / 2$.

<u>Step 3</u>. By what we know from Unit 8.4 the confidence interval is

$$P(\bar{d} - ME < \mu_d < \bar{d} + ME) = 1 - \alpha, \ \ ME = t_{n-1,\alpha/2} \frac{s_d}{\sqrt{n}}.$$

It remains to replace μ_d by $\mu_X - \mu_Y$.

Unit 9.2   Independent samples. Case I: σ_X, σ_Y known

When efficiency of a new medicine is tested, usually its effect on a group of patients is compared to the health of members of a control group, who are not administered any

medication. In this case we have two distinct samples of different sizes n_X, n_Y. They come independently from two different populations.

Assumptions. The samples $x = (x_1, ..., x_{n_X})$, $y = (y_1, ..., y_{n_Y})$ are independent normal. The population means are unknown and the population variances are known.

By properties of means and variances

$$E(\bar{X} - \bar{Y}) = E\bar{X} - E\bar{Y} = \mu_X - \mu_Y, \ V(\bar{X} - \bar{Y}) = V(\bar{X}) + V(\bar{Y}) = \frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}. \quad (9.1)$$

Finding the confidence interval

Step 1. Find the sample means.

Step 2. Choose the level of significance α and find the critical value of the z statistic from $P(z_{n-1} > z_{\alpha/2}) = \alpha / 2$.

Step 3. Then the confidence interval for $\mu_X - \mu_Y$ is

$$P(\bar{X} - \bar{Y} - ME < \mu_X - \mu_Y < \bar{X} - \bar{Y} + ME) = 1 - \alpha, \ ME = z_{\alpha/2} \sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}.$$

Unit 9.3　Independent samples. Case II: σ_X, σ_Y unknown but equal (pooled estimator)

When $\sigma_X^2 = \sigma_Y^2 = \sigma^2$ from (9.1) we see that

$$\sigma(\bar{X} - \bar{Y}) = \sigma \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}. \quad (9.2)$$

The question therefore is how one combines the information contained in two samples to better estimate σ? **Theorem 7.1** provides the leading idea in this and the next section.

Idea. The population variance σ^2, sample variance s^2 and degrees of freedom r should satisfy the equation

$$\chi_r^2 = \frac{rs^2}{\sigma^2} \quad (9.3)$$

To find the right estimator we use an inductive argument. By **Theorem 7.1** we have equations

$$\chi_{n_X-1}^2 = \frac{(n_X - 1)s_X^2}{\sigma^2}, \ \chi_{n_Y-1}^2 = \frac{(n_Y - 1)s_Y^2}{\sigma^2}.$$

Summing them and using the fact that the two chi-square distributions involved are independent we get

$$\frac{(n_X - 1)s_X^2}{\sigma^2} + \frac{(n_Y - 1)s_Y^2}{\sigma^2} = \chi_{n_X-1}^2 + \chi_{n_Y-1}^2 = \chi_{n_X+n_Y-2}^2. \quad (9.4)$$

From (9.3) and (9.4) we see that the estimator of σ^2 should be defined by

$$s_p^2 = \frac{(n_X - 1)s_X^2 + (n_Y - 1)s_Y^2}{n_X + n_Y - 2}. \tag{9.5}$$

Definition. (9.5) is called a *pooled estimator*.

Exercise 9.1 (1) In addition to (9.3), the pooled estimator has the usual unbiasedness property $Es_p^2 = \sigma^2$.

(2) Besides, it gives more weight to the sample with the larger size.

(3) Obviously, (9.2) is estimated by $\sigma(\bar{X} - \bar{Y}) = s_p \sqrt{\dfrac{1}{n_X} + \dfrac{1}{n_Y}}$. Following the general statistical idea, in

$$z_0 = \frac{\bar{X} - \bar{Y}}{\sigma(\bar{X} - \bar{Y})} = \frac{\bar{X} - \bar{Y}}{\sigma \sqrt{\dfrac{1}{n_X} + \dfrac{1}{n_Y}}}$$

replace σ by the pooled estimator to obtain

$$t = \frac{\bar{X} - \bar{Y}}{s_p \sqrt{\dfrac{1}{n_X} + \dfrac{1}{n_Y}}}$$

Show that this t is distributed as $t_{n_X + n_Y - 2}$.

Finding the confidence interval

<u>Step 1</u>. Calculate $\bar{X}, \bar{Y}, s_X^2, s_Y^2, s_p^2$.

<u>Step 2</u>. Choose α and find the critical value of the t statistic $t_{n_X + n_Y - 2, \alpha/2}$.

<u>Step 3</u>. The confidence interval is $P(\bar{X} - \bar{Y} - ME < \mu_X - \mu_Y < \bar{X} - \bar{Y} + ME) = 1 - \alpha$,

where $ME = t_{n_X + n_Y - 2, \alpha/2} s_p \sqrt{\dfrac{1}{n_X} + \dfrac{1}{n_Y}}$.

Unit 9.4 Independent samples. Case III: σ_X, σ_Y unknown and unequal (Satterthwaite's approximation)

Assumptions. There are two samples of sizes n_X, n_Y, respectively, drawn from normal populations with unknown population means. Population variances σ_X^2, σ_Y^2 are unknown and unequal.

This topic shows how complex and long a preamble can be.

Preamble. Equations (9.1) continue to be true. Denote $D = \bar{X} - \bar{Y}$. Because of (9.1), as usual, σ_D^2 is estimated by $s_D^2 = \dfrac{\sigma_X^2}{n_X} + \dfrac{\sigma_Y^2}{n_Y}$ and instead of $z_0 = \dfrac{D - \mu_D}{\sigma_D} = \dfrac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sigma_D}$ we use

$$t = \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{s_D}. \tag{9.6}$$

The idea used in Case II now sounds like this: find the number of degrees of freedom r in such a way that

$$\chi_r^2 = \frac{r s_D^2}{\sigma_D^2}. \tag{9.7}$$

Suppose (9.7) is true. Taking variance of both sides and using homogeneity we get

$2r = V(\chi_r^2) = \left(\dfrac{r}{\sigma_D^2}\right)^2 V(s_D^2)$ or, rearranging,

$$r = \frac{2\sigma_D^4}{V(s_D^2)}. \tag{9.8}$$

To find $V(s_D^2)$ we use the result from **Exercise 7.3**:

$$V(s_D^2) = \frac{1}{n_X^2} V(s_X^2) + \frac{1}{n_Y^2} V(s_Y^2) = \frac{2\sigma_X^4}{n_X - 1} + \frac{2\sigma_Y^4}{n_Y - 1}.$$

Combining this with (9.8) yields

$$r = \frac{\sigma_D^4}{\dfrac{1}{n_X - 1}\left(\dfrac{\sigma_X^2}{n_X}\right)^2 + \dfrac{1}{n_Y - 1}\left(\dfrac{\sigma_Y^2}{n_Y}\right)^2}. \tag{9.9}$$

Definition. Replacing in (9.9) population variances by their sample counterparts we obtain *Satterthwaite's approximation* for the number of degrees of freedom:

$$v = \frac{\left(\dfrac{s_X^2}{n_X} + \dfrac{s_Y^2}{n_Y}\right)^2}{\dfrac{1}{n_X - 1}\left(\dfrac{s_X^2}{n_X}\right)^2 + \dfrac{1}{n_Y - 1}\left(\dfrac{s_Y^2}{n_Y}\right)^2}.$$

Finding the confidence interval

<u>Step 1</u>. Calculate $\bar{X}, \bar{Y}, s_X^2, s_Y^2, v$.

<u>Step 2</u>. Choose α and find the critical value of the t statistic $t_{v, \alpha/2}$.

<u>Step 3</u>. The confidence interval is $P(\bar{X} - \bar{Y} - ME < \mu_X - \mu_Y < \bar{X} - \bar{Y} + ME) = 1 - \alpha$,

where $ME = t_{v, \alpha/2} \sqrt{\dfrac{s_X^2}{n_X} + \dfrac{s_Y^2}{n_Y}}$.

Unit 9.5   Confidence interval for the difference between two population proportions

Exercise 9.2 Suppose that a random sample of n_X observations from a population with proportion p_X of successes yields a sample proportion $\hat{p}_X$ and that an independent random sample of n_Y observations from a population with proportion p_Y of successes produces a sample proportion $\hat{p}_Y$. Which of the Cases I-III would you apply?

Unit 9.6   Questions for repetition

1. Assuming that the population variances are known, derive the confidence interval for the difference between two means in case of (a) matched pairs and (b) independent samples.

2. In the context of two independent samples with unknown but equal variances: (a) calculate $V(\bar{X} - \bar{Y})$, (b) define the pooled estimator and find its mean and (c) derive the confidence interval.

3. Give the decision tree for testing the difference between two means showing your choices when: (a) σ is known/unknown, (b) the samples are similar (matched) or independent, (c) in case of unknown σ, equal or unequal variances and (d) the parent population is general or Bernoulli. In all cases give the expression for ME.

4. Minimum required: 9.4, 9.6, 9.11, 9.12, 9.18, 9.27, 9.35.

Chapter 10 Hypothesis Testing

Unit 10.1 Concepts of hypothesis testing (null and alternative hypotheses, Type I and Type II errors, Cobb-Douglas production function; increasing, constant and decreasing returns to scale)

In hypothesis testing we assume two possibilities, called *hypotheses*, about the state of nature. They must be mutually exclusive and collectively exhaustive events, like "the accused is guilty" and "the accused is innocent". In practice, the true state of nature is never known, and we have to use statistical evidence to make judgment.

One hypothesis is called a *null hypothesis* and denoted H_0 and the other is called an *alternative hypothesis* and denoted H_a. Rejecting the null, when it is true, we commit an error called *Type I error*. Rejecting the alternative, when it is true, leads to another error, called *Type II error*.

Table 10.1 States of nature and our decisions

		States of nature	
		H_0 is true	H_a is true
Decision	Accept H_0	Right decision	Type II error
	Reject H_0	Type I error	Right decision

The choice of the hypotheses is up to us and depends on the circumstances. If we seek a strong evidence in favor of a particular outcome, that outcome should be designated as the alternative hypothesis, because, as comparison of Unit 10.3 and Unit 10.5 shows, it's easier to control the probability of Type I error.

Example 10.1 In a criminal trial the prosecutor wants to prove that the defendant is guilty, and that would be the alternative. For the defendant's lawyer the alternative is that the defendant is innocent.

Example 10.2 A *Cobb-Douglas production function* is defined by $f(K,L) = AK^\alpha L^\beta$ where A, α, β are constants and K, L are independent variables (capital and labor). It is homogeneous of degree $\alpha + \beta$: $f(tK, tL) = A(tK)^\alpha (tL)^\beta = t^{\alpha+b} f(K,L)$. We say that this function exhibits *increasing*, *constant* or *decreasing returns to scale* if, respectively, $\alpha + \beta > 1$, $\alpha + \beta = 1$ or $\alpha + \beta < 1$. In applied research we might want to test

$$H_0 : \alpha + \beta = 1 \text{ against } H_a : \alpha + \beta \neq 1. \tag{10.1}$$

Example 10.3 We believe that usually the supply curve is positively sloping. If we model it as a straight line, $p = a + bq$, then to prove that it is really positively sloped we test

$$H_0 : b \leq 0 \text{ against } H_a : b > 0. \tag{10.2}$$

Example 10.4 The hypotheses in (10.1) and (10.2) are complementary. In NCT you can see examples of noncomplementary hypotheses, such as

$$H_0 : \alpha + \beta = 1 \text{ against } H_a : \alpha + \beta > 1. \tag{10.3}$$

The same testing procedures are applied in case of noncomplementary hypotheses, just because the statistical science is not perfect.

Unit 10.2 Tradeoff between Type I and Type II errors (significance level, power of a test)

We always try to minimize the probability of error. The probability $\alpha = P(\text{Type I error})$ is called a *significance level*. The probability $\beta = P(\text{Type II error})$ doesn't have a special name but $1 - \beta = P(\text{Accepting } H_a \text{ when it is true})$ is called a *power of a test*. In hypothesis testing we want low α and high $1 - \beta$. Unfortunately, these are contradicting requirements.

Exercise 10.1 Suppose applicants to a university are given a test which is graded on a 100 point scale. The admission committee has to choose the passing grade. In this context discuss the tradeoff between Type I and Type II errors.

Solution. We categorize applicants as "weak" and "strong". These characteristics are not directly observable and the committee has to use the grades as a proxy. It is imperfect: strong applicants may get lower grades than weak ones, due to a number of reasons. In statistics all statements are probabilistic. Committee's task is to make the decision using conditional probabilities

$$p_w = P(\text{applicant is weak} \mid grade), \ p_s = P(\text{applicant is strong} \mid grade).$$

The first of these probabilities decreases with *grade* and the second increases, see Figure 10.1.
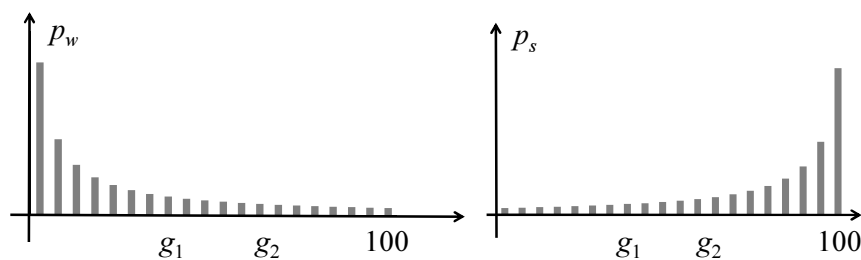
Figure 10.1 Tradeoff between two types of errors

The cut-off level g_2 is better to sort out weak students. The passing grade g_1 is preferable if the objective is to admit all potentially good students. Since these are contradictory objectives, the role of hypotheses is not symmetric: H_a is given preference. If the emphasis is on admission of strong students, the choice is

$$H_0 : \text{applicant is weak}, \ H_a: \text{applicant is strong}.$$

Universities which maximize tuition do the opposite: they don't want to lose anybody who is able to study and pay.

Important observation: the distribution used for decision making is determined by the null hypothesis, see **Exercise 10.2**.

Cautious terminology. When we reject the null, we indicate the significance level to emphasize that our decision may be wrong with probability α. Instead of saying "we accept the null" we say "fail to reject the null" bearing in mind that H_0 can still be wrong.

Unit 10.3   Using confidence intervals for hypothesis testing (decision rule, acceptance and rejection regions, simple and composite hypotheses)

Exercise 10.2 A cereal package should weigh w_0. If it weighs less, the consumers will be unhappy and the company may face legal charges. If it weighs more, company's profits will fall. Thus for the quality control department of the company the right choice is $H_0 : \mu_w = w_0$, $H_a : \mu_w \neq w_0$. Devise the hypothesis testing procedure.

Solution. Step 1. Suppose that the package weight is normally distributed, σ_w is known and $\bar{w}$ is the sample mean based on a sample of size n. From Unit 8.3 we know that

$$P\left(\mu_w - ME < \bar{w} < \mu_w + ME\right) = 1 - \alpha \quad \text{where} \quad ME = z_{\alpha/2} \frac{\sigma_w}{\sqrt{n}}. \qquad (10.4)$$

Under H_0 the confidence interval is $(w_0 - ME < \bar{w} < w_0 + ME)$.

Step 2. Formulate the consequences of the null hypothesis.

(a) If $\mu_w = w_0$, then $\bar{w}$ is likely to be close to w_0. This is expressed by $P(w_0 - ME < \bar{w} < w_0 + ME) = 1 - \alpha$.

(b) On the other hand, under the null the sample mean is not likely to be far from from the assumed mean: $P(|\bar{w} - w_0| \geq ME) = \alpha$.

Step 3. State the *decision rule*.

(a) Fail to reject the null if the sample mean is consistent with it, that is $w_0 - ME < \bar{w} < w_0 + ME$ (this defines the *acceptance region*)

(b) Reject the null if the sample mean is consistent with the alternative: $|\bar{w} - w_0| \geq ME$ (that is, either $\bar{w} \geq w_0 + MEP$ or $\bar{w} \leq w_0 - ME$; this defines the *rejection region*). Then $P(\text{Type I error}) = \alpha$.

The decision rule is illustrated in Figure 10.2.
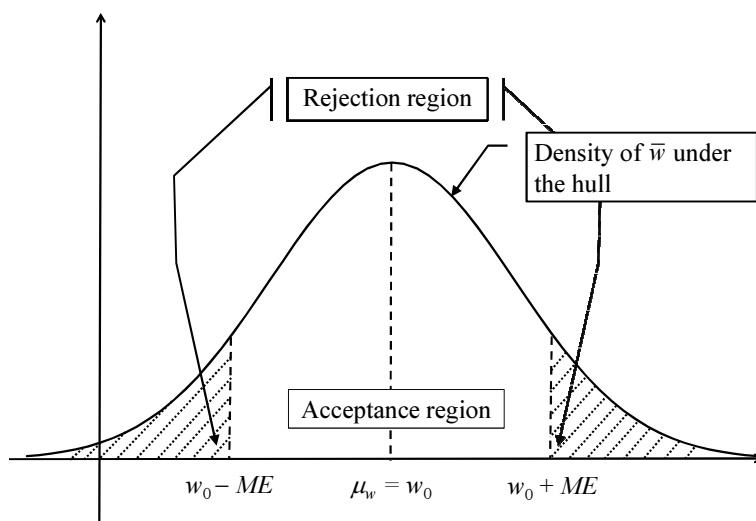
Figure 10.2 Rejection and acceptance regions

Comments. **(1)** When solving exercises you can usually jump to the decision rule but remember that it is always based on confidence intervals. When in doubt, do the derivation and draw the diagram.

(2) The above decision rule is in terms of the sample mean. Equivalently, it can be expressed in terms of the standardized sample mean. When the population is normal and σ_w is known, we use the z statistic $z = \dfrac{\overline{w} - w_0}{\sigma_w / \sqrt{n}}$ and the decision rule looks like this: reject the null if $|z| \geq z_{\alpha/2}$ and fail to reject the null if $|z| < z_{\alpha/2}$.

(3) In theory it doesn't matter if the endpoints of the confidence interval are included in the rejection or acceptance region (the probability of one point is zero). In practice, with the view to smoothly transit to p-values it is better to include the endpoints in the rejection area.

(4) To obtain a confidence interval, we need to fix the distribution by selecting a parameter which satisfies the null. If the null is *simple* (contains only one parameter value), this is easy. If the null is *composite* (contains a whole interval), the parameter is usually fixed at the boundary of the interval. No better procedures have been designed so far.

Unit 10.4   p-values

Preamble. Consider the situation in Figure 10.3.
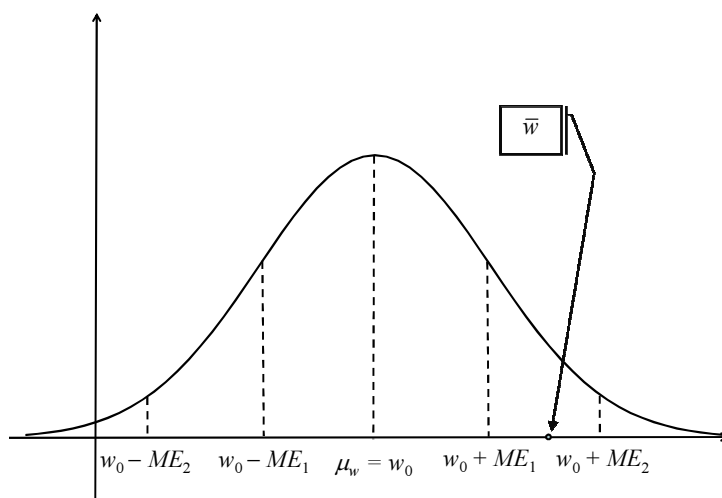
Figure 10.3 Changing level of significance

Let us think about the dynamics of the confidence interval when α changes. When α decreases, the confidence interval becomes wider. The sample mean, being random, can take values both inside and outside a given confidence interval. It is possible to have the situation depicted in Figure 10.3: $w_0 + ME_1 < \overline{w} < w_0 + ME_2$ where ME_i corresponds to α_i, $i = 1, 2$, and $\alpha_2 < \alpha_1$. We can reject the null at the significance level α_1 and would be able to do that at a slightly lower significance level. Similarly, we fail to reject the null at the level of significance α_2 and the decision would be the same at a slightly higher significance level. As it is desirable to reduce the probability of Type I error, this leeway is not good. We would like to know the precise value of the significance level, beyond which we would always fail to reject the null hypothesis.

Definition. The p-*value* is the smallest significance level at which the null hypothesis can still be rejected.

Comments. **(1)** In case of long verbal definitions divide and repeat them in short logical groups:

The p-*value*

                      is the smallest significance level

                             at which the null hypothesis

                                   can still be rejected.

(2) The critical value $z_{\alpha/2}$ in (10.4) is a function of α and is defined from $P(z > z_{\alpha/2}) = \alpha/2$. It doesn't depend on the realized value $\overline{w}$. The p-value, on the other hand, is defined by $P(z > \overline{w}) = p/2$ and depends on the realized sample mean.

(3) It must be clear from the definition that

        (i) the null is rejected at any $\alpha \geq p$ and

        (ii) for any $\alpha < p$ the decision is "Fail to reject the null".

Unit 10.5   Power of a test

There are not many exercises for assessing the power of a test so this will probably be your first and last chance to learn the topic.

From the definition

$$\beta = P(\text{Rejecting the alternative when it is true})$$

we see that we have to find the probability of the *nonrejection region* (= acceptance region), where the nonrejection region is found under the null and the probability is found under the alternative hypothesis, and then put Power = $1 - \beta$. More precisely, there are three steps:

<u>Step 1</u>. Fix the sampling distribution by selecting a parameter that satisfies H_0, choose α and find the nonrejection region N.

<u>Step 2</u>. Suppose that the alternative is true. Fix any value $\mu = \mu^*$ of the parameter that satisfies H_a and find the probability of accepting H_0 under the NEW sampling distribution

$$\beta = P(\text{sample statistic} \in N \mid \mu = \mu^*).$$

β estimates $P(\text{Type II error})$.

<u>Step 3</u>. Then Power = $1 - \beta$.

Example 10.5 For the sample mean $\overline{X}$ test $H_0 : \mu = \mu_0$ against $H_a : \mu > \mu_0$, assuming σ is known and the sample size is n.

<u>Step 1</u>. This is a one-tail test. The margin of error is $ME = z_\alpha \sigma/\sqrt{n}$, the rejection region is $R = \{\overline{X} \geq \mu_0 + ME\}$ and the nonrejection region is $N = \{\overline{X} < \mu_0 + ME\}$, see Figure 10.4.
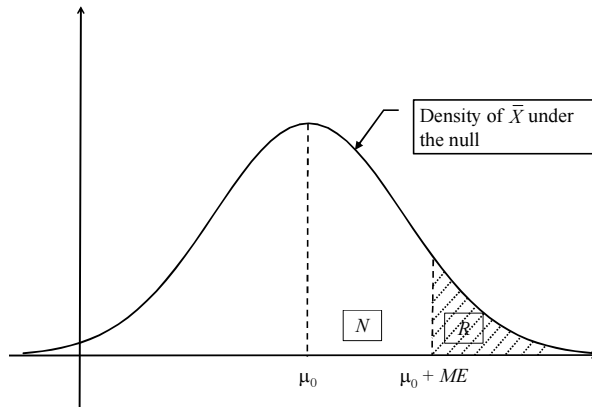
Figure 10.4 The nonrejection region under the null

N has been found under the OLD density $p_{\bar{X}}(t \mid \mu = \mu_0)$.

<u>Step 2</u>. Now instead of μ_0 satisfying H_0 we take μ^* satisfying H_a. With the new parameter we have $E\bar{X} = \mu^*$ and the standardized sample mean is $z = \dfrac{\bar{X} - \mu^*}{\sigma / \sqrt{n}}$. The probability of the OLD region N under the NEW density $p_{\bar{X}}(t \mid \mu = \mu^*)$ is

$$\beta = P(N \mid \mu = \mu^*) = P(\bar{X} \leq \mu_0 + ME \mid \mu = \mu^*)$$

$$= P\left(\frac{\bar{X} - \mu^*}{\sigma / \sqrt{n}} \leq \frac{\mu_0 + ME - \mu^*}{\sigma / \sqrt{n}}\right) = P\left(z \leq \frac{\mu_0 + ME - \mu^*}{\sigma / \sqrt{n}}\right).$$

<u>Step 3, concluding</u>. Power $= 1 - \beta$.

Exercise 10.3 Repeat the procedure for finding the power in one of the following cases:

(a) $H_0 : \mu = \mu_0$, $H_a : \mu > \mu_0$

(b) $H_0 : \mu \leq \mu_0$, $H_a : \mu > \mu_0$

(c) $H_0 : \mu = \mu_0$, $H_a : \mu < \mu_0$

(d) $H_0 : \mu \geq \mu_0$, $H_a : \mu < \mu_0$

(e) $H_0 : \mu = \mu_0$, $H_a : \mu \neq \mu_0$

Answer the following questions:

(i) What happens to the power when μ^* moves away from μ_0?

(ii) What happens to $1 - \beta$ when α decreases?

(iii) How does the sample size affect the power?

(iv) How does $1 - \beta$ behave if σ increases?

(v) Summarize your answers in an intuitive, easy-to-remember form that would not appeal to graphs.

Unit 10.6   Questions for repetition

1. In the context of a university admission test discuss the trade-off between Type I and Type II errors.

2. Explain the statistical logic behind the decision rule using the example of a cereal package (the choice of hypotheses and the use of the confidence interval).

3. z statistics versus p-values

(a) Give economic examples of the null and alternative hypotheses which lead to (i) a one-tail test and (ii) a two-tail test.

(b) How are the critical values defined and what are the decision rules in these cases using z statistics?

(c) Define p values for each of those cases and explain the benefits/drawbacks of using z statistics versus p-values.

4. Normal population versus Bernoulli

(a) Select two pairs of hypotheses such that in one case you would apply a two-tail test and in the other – a one-tail test. Assuming that the population is normal and the variance is unknown, illustrate graphically the procedure of testing the mean.

(b) What changes would you make if the population were Bernoulli?

5. Using the example of testing the sample mean describe in full the procedure of assessing the power of a test. Answer the following questions:

(a) What happens to the power when μ^* moves away from μ_0?

(b) What happens to $1-\beta$ when α decreases?

(c) How does the sample size affect the power?

(d) How does $1-\beta$ behave if σ increases?

6. Give the decision tree for testing the difference between two population means distinguishing between: (a) dependent and independent samples, (b) cases of known and unknown variances, (c) equal and unequal variances when they are unknown and (d) normal and Bernoulli populations.

7. Minimum required: 10.06, 10.13, 10.16, 10.32.

Chapter 11 Hypothesis Testing II

If you have been following the material, as my students do, this chapter must be easy for you. That's why the chapter is short.

Unit 11.1 Testing for equality of two sample variances (F-distribution)

Definition. Let χ^2_m and χ^2_n be independent. Then

$$F_{m,n} = \frac{\chi^2_m / m}{\chi^2_n / n}$$

is called and F -*distribution* with (m, n) degrees of freedom.

Comments. **(1)** Equivalently, F can be defined by

$$F_{m,n} = \frac{\sum_{i=1}^{m} z_i^2 / m}{\sum_{j=1}^{n} z_{m+j}^2 / n}$$

where $z_1, ..., z_m, z_{m+1}, ..., z_{m+n}$ are standard normal independent.

(2) From the definitions of t and F it follows that $t_n^2 = F_{1,n}$. This fact is used in simple regression (a t test of significance of a regression coefficient is equivalent to the F test of significance).

(3) We know that for independent samples from two normal populations

$$\frac{(n_X - 1)s_X^2}{\sigma_X^2} = \chi^2_{n_X - 1}, \ \frac{(n_Y - 1)s_Y^2}{\sigma_Y^2} = \chi^2_{n_Y - 1}.$$

Hence,

$$\frac{s_X^2 / \sigma_X^2}{s_Y^2 / \sigma_Y^2} = \frac{\chi^2_{n_X - 1} / (n_X - 1)}{\chi^2_{n_Y - 1} / (n_Y - 1)} = F_{n_X - 1, n_Y - 1}.$$

(4) To test the null $H_0 : \sigma_X^2 = \sigma_Y^2$ against the alternative $H_0 : \sigma_X^2 \neq \sigma_Y^2$ choose α, find the critical values for the two-tail test from

$$P(F > F_{n_X - 1, n_Y - 1, 1 - \alpha/2}) = 1 - \alpha/2, \ P(F > F_{n_X - 1, n_Y - 1, \alpha/2}) = \alpha/2$$

(the notation of critical values follows Table 9 from NCT). Then under the null

$$1 = \alpha = P(F_{n_X - 1, n_Y - 1, 1 - \alpha/2} < F < F_{n_X - 1, n_Y - 1, \alpha/2})$$

$$= P\left(F_{n_X - 1, n_Y - 1, 1 - \alpha/2} < \frac{s_X^2}{s_Y^2} < F_{n_X - 1, n_Y - 1, \alpha/2}\right).$$

The null is rejected if s_X^2 / s_Y^2 either exceeds the upper critical value or is smaller than the lower critical value.

Unit 11.2   Questions for repetition

1. In one block define and describe properties of chi-square, t and F distributions.

2. Testing for equality of variances

   (a) What is the distribution of the ratio of two sample variances from two normal populations?

   (b) How would you test for equality of variances from two normal populations?

2. Minimum required: 11.24, 11.28.

Chapter 12   Linear Regression

Unlike the previous chapter, this one will be densely populated. It's because solutions to some exercises from NCT are completely spelled out. This chapter covers the first chapter of a standard econometrics text.

Unit 12.1   Algebra of sample means

For a discrete random variable X with $x_1,...,x_n$ and uniform distribution denote $E_u X = x_1 \frac{1}{n} + ... + x_n \frac{1}{n}$. This is just the sample mean: $E_u X = \bar{X}$. Therefore all properties we've established for the mean value operator translate to corresponding properties of sample means. We need:

(a) linearity of means $\overline{(aX + bY)} = a\bar{X} + b\bar{Y}$,

(b) alternative expression for covariance $\overline{(X - \bar{X})(Y - \bar{Y})} = \overline{XY} - \bar{X}\bar{Y}$ and

(c) variance $\overline{(X - \bar{X})^2} = \overline{X^2} - (\bar{X})^2$.

To see usefulness of these abbreviations, compare the last equation to its full form:

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n}\sum_{i=1}^{n}X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)^2.$$

The tricky part of the notation just introduced is that $X_1,...,X_n$ can be treated not as real values of a random variable X but as random variables sampled from the same population (this is why they are written using capital X's). To emphasize the distinction between the population mean and sample mean, I use the notation E_u for the latter. It give rise to $\text{cov}_u(X,Y) = \overline{(X - \bar{X})(Y - \bar{Y})}$ and $V_u(X) = \overline{(X - \bar{X})^2}$.

Unit 12.2   Linear model setup (linear model, error term, explanatory variable, regressor)

We assume that vectors (x_1, y_1), ..., (x_n, y_n) are observed and want to capture dependence of y_i on x_i. The *linear model* seeks a linear dependence $y_i = \beta_1 + \beta_2 x_i$. Since usually data are chaotic, exact linear dependence is impossible and it is more realistic to suppose that y_i differs from a linear function $\beta_1 + \beta_2 x_i$ by a random *error term* u_i. Thus, the main assumption is

$$y_i = \beta_1 + \beta_2 x_i + u_i, \ i = 1,...,n. \tag{12.1}$$

The independent variable x is called an *explanatory variable* or a *regressor*.

Regarding the random variables $u_1,...,u_n$ for now we assume that their mean is zero. The regressor is assumed to be deterministic.

An immediate consequence of this is that the dependence of population means of y_i on x_i is linear: $Ey_i = \beta_1 + \beta_2 x_i$. Out first task is to estimate the intercept β_1 and the slope β_2.

Unit 12.3   Ordinary least squares estimation (OLS estimators, normal equations, working formula)

Definition (Gauss). Follow this definition on Figure 12.1. Take an arbitrary straight line $y = b_1 + b_2 x$, find the residuals $y_i - (b_1 + b_2 x_i)$ from approximation of the observed values y_i by the corresponding values $b_1 + b_2 x_i$ along the straight line and accept as *OLS estimators* of β_1 and β_2 those values b_1, b_2 which minimize the sum of squared residuals

$$f(b_1, b_2) = \sum_{i=1}^{n} (y_i - b_1 - b_2 x_i)^2.$$
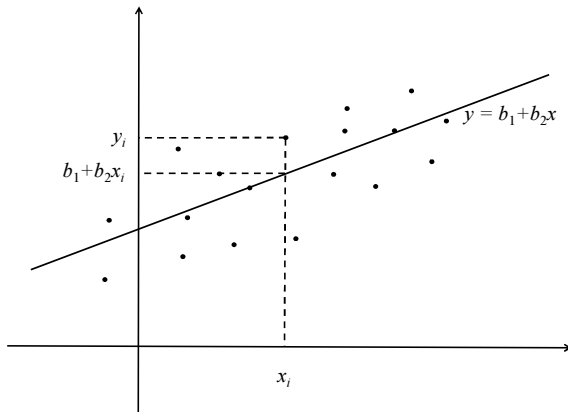
Figure 12.1 Illustration of the OLS procedure

12.3.1   Derivation of OLS estimators

The first order conditions for the minimization of $f(b_1, b_2)$ are

$$\frac{\partial f(b_1, b_2)}{\partial b_1} = -2 \sum_{i=1}^{n} (y_i - b_1 - b_2 x_i) = 0,$$

$$\frac{\partial f(b_1, b_2)}{\partial b_2} = -2 \sum_{i=1}^{n} (y_i - b_1 - b_2 x_i) x_i = 0.$$

Getting rid of -2 and dividing both equations by n we get

$$\frac{1}{n} \sum_{i=1}^{n} (y_i - b_1 - b_2 x_i) = 0, \quad \frac{1}{n} \sum_{i=1}^{n} (y_i - b_1 - b_2 x_i) x_i = 0. \tag{12.2}$$

Introducing the vector notation

$$y = \begin{pmatrix} y_1 \\ ... \\ y_n \end{pmatrix}, \ x = \begin{pmatrix} x_1 \\ ... \\ x_n \end{pmatrix}, \ u = \begin{pmatrix} u_1 \\ ... \\ u_n \end{pmatrix} \tag{12.3}$$

we can write (12.2) as $\overline{y - b_1 - b_2 x} = 0, \ \overline{(y - b_1 - b_2 x)x} = 0.$ Using linearity of sample means rearrange this:

$$\overline{y} - b_1 - b_2 \overline{x} = 0, \ \overline{yx} - b_1 \overline{x} - b_2 \overline{x^2} = 0. \tag{12.4}$$

We have obtained what is called *normal equations*.

From the first of them we see that we can find

$$b_1 = \bar{y} - b_2 \bar{x} \tag{12.5}$$

if we know b_2. To find b_2, plug (12.5) into the second normal equation:

$$\overline{yx} - (\bar{x})(\bar{y}) + b_2 \bar{x}^2 - b_2 \overline{x^2} = 0.$$

This last equation implies

$$b_2 = \frac{\overline{yx} - (\bar{x})(\bar{y})}{\overline{x^2} - \bar{x}^2} = \frac{\dfrac{1}{n}\sum_{i=1}^{n} x_i y_i - \left(\dfrac{1}{n}\sum_{i=1}^{n} x_i\right)\left(\dfrac{1}{n}\sum_{i=1}^{n} y_i\right)}{\dfrac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\dfrac{1}{n}\sum_{i=1}^{n} x_i\right)^2} = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}. \tag{12.6}$$

(12.5) and (12.6) are called *OLS estimators*. Applying notation from Unit 12.1 we get an equivalent expression for the estimator of the slope:

$$b_2 = \frac{\mathrm{cov}_u(x, y)}{V_u(x)}. \tag{12.7}$$

The last part of (12.6) is better for calculations. (12.7) is better for establishing theoretical properties.

12.3.2 Unbiasedness of OLS estimators

Exercise 12.1 (1) If $V_u(x) \neq 0$ then the OLS estimators exist.

(2) If, further, $Eu_i = 0$ for all i and x_i are deterministic , then the OLS estimators are unbiased.

Proof. The statement about existence is obvious. Using notation (12.3) let us write the system of equations (12.1) in a vector form

$$y = \beta_1 + \beta_2 x + u.$$

Using linearity of covariances we rearrange the slope estimator

$$b_2 = \frac{\mathrm{cov}_u(x, \beta_1 + \beta_2 x + u)}{V_u(x)} = \frac{\mathrm{cov}_u(x, \beta_1) + \beta_2\,\mathrm{cov}_u(x, x) + \mathrm{cov}_u(x, u)}{V_u(x)}.$$

Covariance of a constant with anything is zero. Covariance of a variable with itself is its variance. Therefore from the last equation we obtain a *working formula* for the slope estimator:

$$b_2 = \beta_2 + \frac{\mathrm{cov}_u(x, u)}{V_u(x)}. \tag{12.8}$$

Since, by assumption, x is deterministic and the error has mean zero, we get unbiasedness of the slope estimator

$$Eb_2 = \beta_2 + \frac{\mathrm{cov}_u(x, Eu)}{V_u(x)} = \beta_2.$$

The vector form of the linear model implies

$$\bar{y} = \beta_1 + \beta_2 \bar{x} + \bar{u}.$$

We use the last two equations to prove unbiasedness of the intercept estimator:

$$Eb_1 = E(\beta_1 + \beta_2\bar{x} + \bar{u} - b_2\bar{x}) = \beta_1 + \beta_2\bar{x} + \underbrace{E\bar{u}}_{=0} - \underbrace{(Eb_2)}_{=\beta_2}\bar{x} = \beta_1.$$

Unit 12.4 Variances of OLS estimators (homoscedasticity, autocorrelation, standard errors)

In addition to the assumptions of **Exercise 12.1** suppose that

$$Eu_1^2 = ... = Eu_n^2 = \sigma_u^2 \text{ (homoscedasticity)} \qquad (12.9)$$

$$Eu_iu_j = 0 \text{ for all } i \neq j \text{ (absence of autocorrelation)}. \qquad (12.10)$$

Comments. **(1)** *Homoscedasticity* means that variances of the error terms are the same across observations. It holds when the errors are identically distributed. In economics usually variances of variables are proportional to their levels (variation in income grows with income) so the errors are often heteroscedastic.

(2) *Absence of autocorrelation* holds for independent errors (provided that their means are null). The condition can be equivalently written as $\mathrm{cov}(u_i, u_j) = 0$, $i \neq j$. Time series data are usually autocorrelated.

Exercise 12.2 Prove that under the above assumptions

$$\sigma_{b_2}^2 = \frac{\sigma_u^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \ \sigma_{b_1}^2 = \sigma_u^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right). \qquad (12.11)$$

Proof. <u>Step 1</u>. Suppose that a random variable Z is a linear combination of the errors with constant coefficients:

$$Z = \sum a_i u_i. \qquad (12.12)$$

Then

$$EZ^2 = \sigma_u^2 \sum a_i^2.$$

To prove this equation it is useful to visualize the multiplication process. When calculating $Z^2 = \left(\sum a_i u_i \right)\left(\sum a_j u_j \right)$ each element in the first pair of parentheses is multiplied by each element in the second pair. Therefore Z^2 is a sum of the products in the table

Table 12.1 Visualizing a squared sum

$a_1u_1a_1u_1$	$\cdots$	$a_1u_1a_nu_n$
$\cdots$	$\cdots$	$\cdots$
$a_nu_na_1u_1$	$\cdots$	$a_nu_na_nu_n$

Separating the diagonal products from off-diagonal and using conditions (12.9)-(12.10) we get

$$EZ^2 = \sum_{i=1}^{n} a_i^2 E u_i^2 + \sum_{i \neq j} a_i a_j E u_i u_j = \sigma_u^2 \sum a_i^2.$$

<u>Step 2</u>. To make use of Step 1, we obtain a representation of type (12.12) for b_2. Rewrite the working formula (12.8) as

$$
\begin{aligned}
b_2 &= \beta_2 + \frac{\text{cov}_u(x,u)}{V_u(x)} = \beta_2 + \frac{\sum_{i=1}^{n}(x_i - \bar{x})(u_i - \bar{u})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\
&= \beta_2 + \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}(u_i - \bar{u}) = \beta_2 + \sum_{i=1}^{n} a_i(u_i - \bar{u})
\end{aligned}
\tag{12.13}
$$

where

$$a_i = \frac{x_i - \bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{x_i - \bar{x}}{n V_u(x)}.$$

The coefficients a_i are deterministic and satisfy

$$\sum_{i=1}^{n} a_i = n\bar{a} = \frac{n}{n V_u(x)} \overline{x - \bar{x}} = 0.$$

This allows us to simplify (12.13)

$$b_2 = \beta_2 + \sum_{i=1}^{n} a_i(u_i - \bar{u}) = \beta_2 + \sum_{i=1}^{n} a_i u_i - \bar{u}\sum_{i=1}^{n} a_i = \beta_2 + \sum_{i=1}^{n} a_i u_i.$$

Therefore by unbiasedness and Step 1

$$\sigma_{b_2}^2 = E(b_2 - Eb_2)^2 = E(b_2 - \beta_2)^2 = E\sum_{i=1}^{n}(a_i x_i)^2 = \sigma_u^2 \sum a_i^2. \tag{12.14}$$

The sum of squares involved here can be found using properties of means

$$\sum_{i=1}^{n} a_i^2 = n\overline{a^2} = \frac{n}{n^2 V_u^2(x)} \overline{(x - \bar{x})^2} = \frac{V_u(x)}{n V_u^2(x)} = \frac{1}{\sum(x_i - \bar{x})^2}.$$

The last two equations prove the first equation in (12.11).

<u>Step 3</u>. We need to derive a similar representation for

$$
\begin{aligned}
b_1 &= \beta_1 + \beta_2\bar{x} + \bar{u} - \left(\beta_2 + \sum a_i u_i\right)\bar{x} \\
&= \beta_1 + \frac{1}{n}\sum u_i - \sum a_i \bar{x} u_i = \beta_1 + \sum\left(\frac{1}{n} - a_i\bar{x}\right)u_i.
\end{aligned}
$$

With the notation $c_i = \frac{1}{n} - a_i\bar{x}$ this becomes $b_1 = \beta_1 + \sum c_i u_i$. In the way similar to (12.14) we get

$$\sigma_{b_1}^2 = E(b_1 - Eb_1)^2 = E(b_1 - \beta_1)^2 = E\sum_{i=1}^{n}(c_i x_i)^2 = \sigma_u^2 \sum c_i^2.$$

111

It remains to calculate the sum of squares

$$\overline{c^2} = \overline{\left(\frac{1}{n} - a\overline{x}\right)^2} = \overline{\left(\frac{1}{n^2} - 2a\overline{x}\frac{1}{n} + a^2\overline{x}\right)}$$

$$= \frac{1}{n^2} - 2\underbrace{(\overline{a})}_{=0}(\overline{x})\frac{1}{n} + \overline{a^2}(\overline{x})^2 = \frac{1}{n^2} + \frac{(\overline{x})^2}{n\sum(x_i - \overline{x})^2}.$$

The second of the formulas (12.11) follows from the last two equations.

12.4.1   Statistics for testing hypotheses

Definition. *Standard errors* of OLS estimators are defined by

$$s_{b_2}^2 = \frac{s_u^2}{\sum\limits_{i=1}^{n}(x_i - \overline{x})^2}, \ s_{b_1}^2 = s_u^2\left(\frac{1}{n} + \frac{\overline{x}^2}{\sum\limits_{i=1}^{n}(x_i - \overline{x})^2}\right).$$

Theorem 12.1 Let the regressor x be deterministic with $V_u(x)$ different from zero. Assume that the random variables $u_1, ..., u_n$ are i.i.d. normal with mean zero and that the true coefficients are β_1^0, β_2^0. Then the variables

$$t_{n-2} = \frac{b_1 - \beta_1^0}{s_{b_1}}, \ t_{n-2} = \frac{b_2 - \beta_2^0}{s_{b_2}}$$

are distributed as t distribution with $n-2$ degrees of freedom.

Unit 12.5   Orthogonality and its consequences (fitted value, residual vector, orthogonality, Pythagorean theorem, Total Sum of Squares, Explained Sum of Squares, Residual Sum of Squares)

After having estimated the coefficients, for prediction purposes we put $\hat{y} = b_1 + b_2 x$. This linear function is called a *fitted value*. It is defined everywhere, both inside and outside the sample. It is customary to use the same notation for the vector $\hat{y} = (b_1 + b_2 x_1, ..., b_1 + b_2 x_n)$ where the fitted value is calculated only at the sample points. Hopefully you will see from the context if the fitted value is a function or a vector. The *residual vector* is defined as $e = y - \hat{y}$.

Definition. Vectors $(x_1, ..., x_n)$, $(y_1, ..., y_n)$ are called *orthogonal* if $x_1 y_1 + ... + x_n y_n = 0$.

Exercise 12.3 (1) Check that the vectors $(0,1)$ and $(1,0)$ are orthogonal.

(2) The notion of uncorrelated random variables has been borrowed from orthogonality. Let X, Y be arbitrary random variables such that $V(X) \neq 0$. Put $Z = Y - \frac{\text{cov}(X,Y)}{V(X)} X$. Prove that then X, Z are uncorrelated. How can you guess a result like this?

First orthogonality relation. $\overline{e} = 0$.

Comments. **(1)** This equation is a reflection of the assumption $Eu = 0$. Notice that in one case the sample mean is used and in the other the population mean.

(2) Keep in mind that for a model without the intercept $y_i = bx_i$ the sum of residuals may not be zero.

(3) This relation also means that the residual is orthogonal to the vector of unities $(1,\dots,1)$:
$\sum e_i = \sum e_i \cdot 1 = n\bar{e} = 0$.

Proof. By the first normal equation (12.4) $\bar{e} = \overline{y - b_1 - b_2 x} = \bar{y} - b_1 - b_2 \bar{x} = 0$.

Corollary 12.1 The fitted value has the same sample mean as y: $\bar{\hat{y}} = \bar{y}$.

Proof. By the definition of the residual vector

$$y = \hat{y} + e \tag{12.15}$$

which implies

$$\bar{y} = \overline{\hat{y} + e} = \bar{\hat{y}} + \bar{e} = \bar{\hat{y}}. \tag{12.16}$$

Second orthogonality relation. The fitted value and residual are orthogonal: $\sum \hat{y}_i e_i = 0$.

Proof. From the second normal equation (12.4)

$$\sum x_i e_i = n\overline{xe} = n\overline{(xy - b_1 x - b_2 x^2)} = n(\overline{xy} - b_1 \bar{x} - b_2 \overline{x^2}).$$

This equation and the first orthogonality relation give

$$\sum \hat{y}_i e_i = \sum (b_1 + b_2 x_i)e_i = b_1 \sum e_i + b_2 \sum x_i e_i = 0.$$

Theorem 12.2 (*Pythagorean theorem*) If the vectors x, y are orthogonal, $\overline{xy} = 0$, then $\overline{(x+y)^2} = \overline{x^2} + \overline{y^2}$.

Proof. $\overline{(x+y)^2} = \overline{x^2 + 2xy + y^2} = \overline{x^2} + 2\overline{xy} + \overline{y^2} = \overline{x^2} + \overline{y^2}$.

Definition. If all y_i are the same, they are all equal to $\bar{y}$ and there is no variation in the dependent variable. Now suppose y_i follows the linear model. The idea is to still measure its deviation from the sample mean $\bar{y}$ but decompose it into two parts: one which is due to the linear part $\beta_1 + \beta_2 x$ and the other which is due to the error. Subtracting the sample mean from both sides of (12.15) we realize this idea as

$$y - \bar{y} = (\hat{y} - \bar{y}) + e. \tag{12.17}$$

This decomposition is accompanied by three definitions:

Total Sum of Squares = measure of magnitude of $y - \bar{y}$: $TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$,

Explained Sum of Squares = measure of magnitude of $\hat{y} - \bar{y}$: $ESS = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$,

Residual Sum of Squares = measure of magnitude of the error: $RSS = \sum_{i=1}^{n} e_i^2$.

Exercise 12.4 $TSS = ESS + RSS$.

Proof. In the representation (12.17) the two terms on the right are orthogonal by the 1st and 2nd orthogonality relations:

$$\overline{(\hat{y} - \overline{y})e} = \overline{\hat{y}e} - \overline{\overline{y}e} = \overline{\hat{y}e} - (\overline{y})(\overline{e}) = 0.$$

By the Pythagorean theorem then

$$TSS = \overline{n(y - \overline{y})^2} = \overline{n(\hat{y} - \overline{y})^2} + \overline{ne^2} = ESS + RSS.$$

Unit 12.6   Goodness of fit (coefficient of determination)

Definition. *The coefficient of determination* is defined by $R^2 = \dfrac{ESS}{TSS}$. It is interpreted as the percentage of total variation of y around the sample mean $\overline{y}$ explained by the regression.

Property 1. From **Exercise 12.4** it follows that

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$$

and that $0 < R^2 < 1$, which justifies its interpretation as percentage.

Property 2. The sample correlation $r_{y,\hat{y}}$ equals $\sqrt{R^2}$.

Proof. By definition and (12.16)

$$r_{y,\hat{y}} = \frac{\text{cov}_u(y,\hat{y})}{\sqrt{V(y)V(\hat{y})}} = \frac{\overline{(y-\overline{y})(\hat{y}-\overline{\hat{y}})}}{\sqrt{\overline{(y-\overline{y})^2}\overline{(\hat{y}-\overline{\hat{y}})^2}}} = \frac{\overline{(y-\overline{y})(\hat{y}-\overline{y})}}{\sqrt{\overline{(y-\overline{y})^2}\overline{(\hat{y}-\overline{y})^2}}}.$$

Here by the orthogonality relations

$$\overline{(y-\overline{y})(\hat{y}-\overline{y})} = \overline{(\hat{y}+e-\overline{y})(\hat{y}-\overline{y})} = \overline{(\hat{y}-\overline{y})^2} + \underbrace{\overline{e(\hat{y}-\overline{y})}}_{=0} = \overline{(\hat{y}-\overline{y})^2}.$$

Therefore

$$r_{y,\hat{y}} = \frac{\overline{(\hat{y}-\overline{y})^2}}{\sqrt{\overline{(y-\overline{y})^2}\overline{(\hat{y}-\overline{y})^2}}} = \sqrt{\frac{\overline{(\hat{y}-\overline{y})^2}}{\overline{(y-\overline{y})^2}}} = \sqrt{R^2}.$$

Unit 12.7   Questions for repetition

1. Explain the setup of simple regression: what is observed, what is the assumed relation between variables, why we have to include the error, the assumptions about the error and how we write everything in a vector form.

2. Write down the formal equations from Unit 12.1 and other properties of means, variances and covariances until you get used to that notation.

3. Define the OLS estimators, derive them and prove unbiasedness. When you do this, do you have to use assumptions (12.9) and (12.10)?

4. Using t statistics, write down confidence intervals for the coefficients. How do you test their significance?

5. How do you evaluate goodness of fit using the coefficient of determination, sum of squared errors and p-values of the estimated coefficients? In other words, if you have two models for the same dependent variable with different regressors, how can you tell which one is better?

6. How are the formulas obtained used for prediction?

7. Minimum required: 12.3, 12.20.

Literature

Michaelsen, Larry K., Knight, Arletta Bauman and Fink, L. Dee. *Team-based Learning: A Transformative Use of Small Groups*. Greenwood Publishing Group. 2002.

Steven G. Krantz. *How to Teach Mathematics*. 2nd ed., American Mathematical Society, Providence.1998.

List of figures

List of tables

Index of terms

116

List of exercises

List of examples

List of theorems