МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ
РОССИЙСКОЙ ФЕДЕРАЦИИ

**Нижегородский государственный университет
им. Н.И. Лобачевского**

А.В. Зорин

Восемь лекций по теории вероятностей и математической статистике

Учебно-методическое пособие

Рекомендовано междисциплинарной методической
комиссией факультета иностранных студентов
для студентов бакалавриата по направлению
38.03.01 «Экономика»

Нижний Новгород
2014

УДК 519.2
ББК Б17
З86

З86 Восемь лекций по теории вероятностей и математической статистике. Автор: Зорин А.В.: Учебно-методическое пособие — Нижний Новгород: Нижегородский госуниверситет, 2014. — 108 с.

Рецензент: к.ф.-м.н., доцент **О.А. Кузенков**

Данное пособие охватывает вводные разделы теории вероятностей и математической статистики. Основное внимание сконцентрировано на случайных величинах и их распределениях вероятностей. Центральная пределеная теорема и закон больших чисел обсуждаются с их практических сторон. Раздел по математичской статистике касается оценки параметров, критериев согласия и простой линейной регрессии.

Учебное пособие предназначено для студентов бакалавриата по направлению «Экономика». Оно также может быть использовано как вспомогателый материал для студентов по направлению «Информационные технологии».

Ответственный за выпуск:
председатель междисциплинарной методической комиссии
факультета иностранных студентов ННГУ,
к.с.н., доцент **Л.В. Ерушкина**

УДК 519.2
ББК Б17

UDC 519.2
BBK B17
Z86

Z86 Eight lectures in probability theory and mathematical statistics. Author: Zorine A.V.: A course boook — Nizhni Novgorod: Lobachevsky State University of Nizhni Novgorod, 2014. — 108 p.

Reviewer: cand.phys.math.sci., docent **O.A. Kuzenkov**

The coursebook covers introductory areas in probability theory and mathematical statistics. Its main focus is on random variables and their probability distributions. The central limit theorem and the law of large numbers are discussed in practical aspects. The mathematical statistics part concerns parameter estimation, goodness-of-fit tests and simple linear regression.

The coursebook is intended for students of the bachelor program in economics. It can also be used as a supplementary reading for students in information technologies.

Supervisor:
the chairman of the Interdisciplinary Methodological Commissions
of the Dept. of International Stidents of NNSU,
cand.soc.sci., docent **L.V. Erushkina**

UDC 519.2
BBK B17

Contents

Foreword

Scientific method relies on measurements of quantitative characteristics of phenomena of study. In economics as well as on other social sciences entities under study are not unified and they differ in their quantitative characteristics in unpredictable way. The differences are usually caused by simultaneous influence of many uncontrolled factors. Nevertheless, a sort of stability can be observed when relative frequencies are taken into account. Probability theory and mathematical statistics were designed to catch and model this type of randomness in many areas of science. Thus the course is intended to prepare students for analytic and researcher activities in economics. A bachelor in Economics is supposed to be able to gather and process statistical data necessary for current economic calculations, as well as be able to develop and test econometric models. By the end of this course, students will be able to develop simple probabilistic models of real-life economic processes, they will be prepared for reading and understanding research reports in stochastic econometrics, prepared to apply probabilistic and statistical methodology to their own research work.

The specifics of this course is a small amount of lecture hours. Hence the Author had to select topics and examples carefully. The preference was given to random variables and to those probability theory results exploited in mathematical statistics. For instance, basic probability theory was explained in terms of events generated by random variables, the classic probability appears in form of uniform joint probability distribution. The examples in this coursebook concern different aspects of economics. Some of the examples take several lectures to develop. Interested reader will find additional practical applications and informative discussion of probability theory and statistics in books [1, 2, 3, 4, 5, 6, 7].

Standard mathematical notations are used throughout the book. A white square □ marks the end of a proof, and symbols ■□■ delimit the end of an example. Numbers in parentheses label equations, e.g. (3.5) means formula number 5 in Lecture 3, numbers in square brackets refer to the list of references in the end of the book.

Lecture 1

Randomness and probability. Basic notions of probability theory

Scientific method is a cornerstone in modern natural and social sciences. Since 17th century, acquiring new knowledge is based on collecting empirical data. As a rule, such data is a result of measurements. One conducts an experiment or observes a natural process paying attention to a selected aspect of an object. Such an *experiment* (in physics, chemistry, biology, etc) or an *observation* (in economics, astronomy, sociology, psychology, etc) must be planned beforehand. In particular, *conditions* should be predefined as precisely as possible. To represent the measurement results scientists use *variables* — quantities that can be given any value from a set of values. For example, a motion of a solid body in space (Fig. 1.1) can be described by six variables: the spatial Cartesian coordinates (x, y, z), and the velocity components (v_x, v_y, v_z). These quantities vary in time, they are functions of a time variable, t. Gross Domestic Product (the overall market values of goods and services produced by labour and capital during one year) is an important measure of an economy of a country (Fig. 1.2).

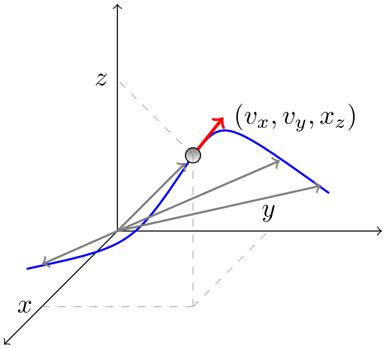

Figure 1.1. Motion of a solid body. A blue curved line is its trajectory. Gray arrows indicate several sequential positions of the body. Dashed lines give its instant Cartesian coordinates. Velocity is depicted by the red thick arrow

In what follows we will consider the notions of *experiments*, *conditions*

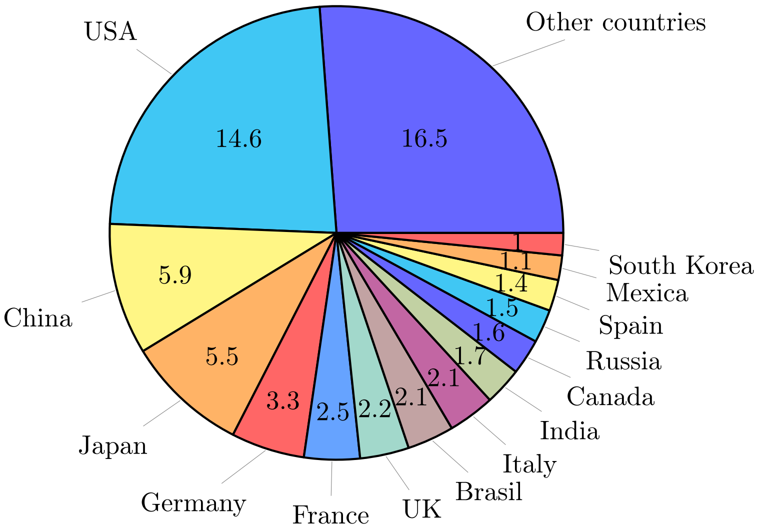

Figure 1.2. Countries by GDP $ trillion, with GDP over $1 trillion in 2010 (*Source:* World Bank Open Data. http://data.worldbank.com)

of experiment, *variables*[1] as elementary and undefinable notions just like *points*, *lines*, and *planes* are undefinable in geometry, or a *mass*, *space*, and *time* in physics. Many sciences use such notions taken from our life experience without strict definition. Such notions are usually explained by many examples.

Now let us choose an interval (a, b) of values of a variable X.

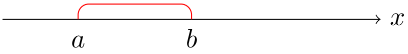

Figure 1.3. A set of values x such that $a < x < b$

After an observation of the variable X has been carried out and the value taken on by X has been determined one can tell whether inequalities

$$a < X < b$$

[1] Strictly speaking, variables have a mathematical definition in more advanced probability theory.

are fulfilled or not. (We might talk about an equality of the kind $X = a$ instead, but we'd like to stress that often we are unable to measure the variable with perfect accuracy because every measuring instrument has its rounding error. Thus the intervals are more illustrative.) It turns out that experiments differ in the following respect: if the inequalities $a < X < b$ were observed previously, they will be always true in the next experiments when the conditions of the experiment are adhered. Such experiments are called *deterministic*. Deterministic experiments led to a notion of a law of nature. This predictability of results given the conditions allowed the usage of mathematics to make such predictions. Many laws in classical physics, chemistry, astronomy, such as the law of gravity, Kepler's laws of planet motion around the Sun, quantitative relations for chemical substances in reactions have this form.

But there are another experiments as simple as rolling dice which don't have such deterministic nature. Let us consider the dice example. Let the variable X count the number on points on the upper face of a die rolled once. It's obvious that X can be any number from the set of 1, 2, 3, 4, 5, and 6. But before rolling the die we can't predict the number of points thrown. Further, if the number of points shown is 1, then the next roll may show any number of points from 1 to 6 anyway. This kind of experiments is called *nondeterministic*, or just *random*. The variable X in this case is called *random variable* as well. (Older names are *chance variable*, *stochastic variable*.) What kind of law can exist when 'everything is random'? When we study mass-scale experiments new regularities appear. Let us assume that the random experiment was executed N times and the variable X was measured each time. Denote by $N(a, b)$ the number of times the inequalities $a < X < b$ were observed. The ratio

$$F_N^*(a, b) = \frac{N(a, b)}{N}$$

is called the *relative frequency* of hitting the interval (a, b). Suppose, this relative frequency varies a little from one series of repetitions to another. Suppose moreover, the relative frequency oscillates in a narrow region as the number N of repetitions increases, i.e. $N \to \infty$. Then the relative frequency is said to be *statistically stable*.

Example 1: Groom's age distribution. The data in Table 1.1 was borrowed from UN demographic statistics[2]. The numbers obviously vary

[2] 2009-2010 Demographic Yearbook. — NY: United Nations, 2011.

from one year to another, although the total number of marriages is approximately the same. Moreover the change in the number of grooms within an age interval can be caused both by the change in the total number of grooms and other random sources. Next let us look at the corresponding relative frequencies (Table 1.2). The numbers for different years are close. Of course, this is an assumption which one better tests by special statistical techniques. So far we may believe that the statistical stability assumption holds for groom ages of Costa-Rica.

Table 1.1. Groom ages in Costa-Rica according to UN data

Year / Age	15–19	20–24	25–29	30–34	35–39	40–44
2009	655	4 786	6 587	4 330	2 374	1 507
2010	520	4 572	6 423	4 663	2 484	1 561
Year / Age	**45–49**	**50–54**	**55–59**	**>60**	**Unknown**	**Total**
2009	1 086	693	470	733	698	23 919
2010	1 099	744	462	800	627	23 955

Table 1.2. Relative frequencies for the groom's ages in Costa-Rica according to UN data

Year / Age	15–19	20–24	25–29	30–34	35–39	40–44
2009	0.0274	0.2001	0.2754	0.1810	0.993	0.630
2010	0.0217	0.1909	0.2681	0.1947	0.1037	0.0652
Year / Age	**45–49**	**50–54**	**55–59**	**>60**	**Unknown**	
2009	0.0454	0.0290	0.0196	0.0306	0.0292	
2010	0.0459	0.0311	0.0193	0.0334	0.0262	

Example 2: Quetelet's Murders data.

A classical example of statistical stability belongs to L. Quetelet[3]. He studied data from French criminal justice on murders. The data is shown in Table 1.3. The variable X under study is qualitative, i.e. it takes on non-numerical values. The values are shown in the first column of the Table. Quetelet wrote, "...In every thing which relates to crimes, the same numbers are reproduced so constantly, that it becomes impossible to misapprehend it — even in respect to those crimes which seem perfectly beyond

[3] Lambert Adolphe Jacques Quetelet [French: kətlɛ] (1796–1874) was a Belgian astronomer, mathematician, statistician, and sociologist. He had introduced and popularized quantitative statistical methods for social sciences.

human foresight… nevertheless, experience proves that murders are committed annually not only pretty nearly to the same extent, but even that the instruments employed are in the same proportion.” So, he stressed the stability of relative frequencies (he called them *proportions*). He also indicated a possible use of this stability: “We might even predict annually how many individuals will stain their hands with the blood of their fellow-men, how many will be forgers, how many will deal in poison, pretty nearly in the same way as we may foretell the annual births and deaths.”

Table 1.3. The results of the reports of criminal justice in France for six years (*Source*: L. Quetelet, Treatise on Man and the Development of his Faculties. — Edinburgh, 1842)

Reason / Year	1826	1827	1828	1829	1830	1831
Guns and pistols	56	64	60	61	57	88
Sabre, sword, etc.	15	7	8	7	12	30
Knife	39	40	34	46	44	34
Cudgel, cane, etc.	23	28	31	24	12	21
Stones	20	20	21	21	11	9
Cutting, stabbing, and bruising instruments	35	40	42	45	46	49
Strangulations	2	5	2	2	2	4
By precipitating and drowning	6	16	6	1	4	3
Kicks and blows with the fist	28	12	21	23	17	26
Fire	–	1	–	1	–	–
Unknown	17	1	2	–	2	2
Murders, in total	241	234	227	231	205	266

Let us introduce one more important term: an *event*. Besides inequalities $a < X < b$ we might be interested in another statements like $X < a$, $X \geqslant b$, $X = a$, $X \neq b$, $|X - a| > b$, etc. Moreover, if we have two variables, X and Y, we might be interested in the following statements: $a < X + Y < b$, $X^2 + Y^2 > b$, $|X| < Y$, etc. The corresponding sets of favorable values of the variables X, Y are the interval $(-\infty, a)$, the semi-closed interval $[b, \infty)$, a point $\{a\}$, a line with a removed point $\{x\colon -\infty < x < \infty, x \neq a\} = (-\infty, b) \cup (b, \infty)$, a union of two open rays $(\infty, a - b) \cup (a + b, \infty)$ (Fig 1.4), a strip $\{(x, y)\colon a < x + y < b\}$ (Fig 1.5,

a)), an outer region $\{(x,y)\colon x^2+y^2>b\}$ of a circle of radius $\sqrt{b}$ centered at the origin (Fig 1.5, b)), and an angle $\{(x,y)\colon |x|<y\}$ (Fig. 1.5, c)).

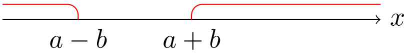

Figure 1.4. A set of values x such that $|x-a|>b$

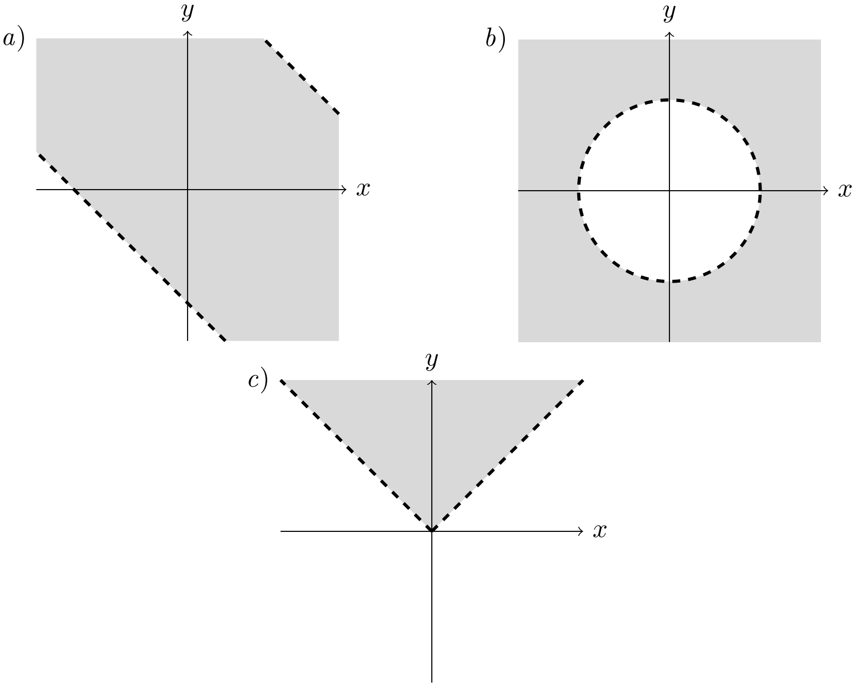

Figure 1.5. Images of the sets: a) $\{(x,y)\colon a<x+y<b\}$, b) $\{(x,y)\colon x^2+y^2>b\}$, c) $\{(x,y)\colon |x|<y\}$

The statements can also be formulated in natural languages (English, Russian, etc.). Here $X^2+Y^2>b$ can be read as “the distance from the origin to the point (X,Y) is greater than $\sqrt{b}$”, and so on.

We will call such statements *events*. We will also call the mathematical sets of points depicting them *events*. We will use different notations for events as statements or inequalities on one hand, and as mathematical sets on the other hand, interchangeably. Generally, we will denote events with capital letters from the beginning of the Latin alphabet, i.e. A, B, etc. Then, same as we did with the random variable X and the interval (a, b), we may compute the relative frequency $F^*_N(G)$ of hitting a set G by the variables of interest. The statistical stability of the experiment means in this case that frequencies $F^*_N(G)$ are close to each other for different series of experiments and their scatterness decreases as N grows.

So, if our previous experience demonstrates that the random experiment is statistically stable we can use previously observed relative frequencies to predict relative frequencies in the future. Moreover, we assume that for given conditions of the random experiment every event has a number assigned to it which represents the *possibility* of occurrence of this event. Such a number is called the *probability of event.* We will denote the probability of event A by $\mathbf{P}(A)$. For instance, equality $\mathbf{P}(A) = p$ means that the numerical value of the probability of event A is p. The frequentists interpretation of probability constitutes that $F^*_N(G) \approx \mathbf{P}(A)$. In other words, a probability is an ideal characteristic of events which can be measured in experiment by means of relative frequency and the measurement typically is not exact.

We have to introduce two important events: Ω the certain event and $\varnothing$ the impossible event. The certain event is the event which is always true because it occurs at each repetition of the experiments, while the impossible is that which is always false because it is never observed. For example, if we have a random variable X that the certain event can be written as "X has taken on one of its values". Notice that if we observe two random variables, X, and Y then two statements: "X has taken on one of its values" and "Y has taken on one of its values" are considered as one and the same certain event. Similarly, the impossible event can be rephrased in many different ways. Rephrasings may have even no relation to the random variables under study, like the statement $-1 > 2$. Nevertheless, we consider all different false statements as a single impossible event. Additional care should be paid when using natural language for events. Surprisingly, "$-1 > 2$, hence the Sun never shines at night" is a true statement. We should also avoid using dubious descriptions of things, objects, numbers, etc., such that "the largest number which can be described with eight words" (to describe this number we used nine words).

The problem of logical foundations of probability theory was clearly formulated by D. Hilbert[4] in 1900. The first set of axioms was proposed

[4] David Hilbert (1862–1943) was a prominent German mathematician. He developed many

by S.N. Bernstein[5] in 1917. He interpreted events as statements following the classical logics style. In his 1933 seminal book A.N. Kolmogorov[6] constructed probability theory on the basis of sets theory and measure theory. Later it was shown that the two approaches are in fact identical.

Mathematical operations on sets, such as the union, the intersection, the difference have direct connection to logical operations on events as Table 1.4 shows. These operations satisfy well-known identities:

$$\begin{gathered}
A \cup \overline{A} = \Omega, \qquad A \cap \overline{A} = \varnothing, \\
A \cap \Omega = A, \quad A \cup \Omega = \Omega, \\
A \cap \varnothing = \varnothing, \quad A \cup \varnothing = A, \\
A \cup (B \cup C) = (A \cup B) \cup C, \\
A \cap (B \cap C) = (A \cap B) \cap C, \\
\overline{A \cap B} = \overline{A} \cup \overline{B}, \qquad \overline{A \cup B} = \overline{A} \cap \overline{B}, \\
A \cap (B \cup C) = (A \cap B) \cup (A \cap C), \\
A \cup (B \cap C) = (A \cup B) \cap (A \cup C).
\end{gathered} \tag{1.1}$$

The most important properties of relative frequencies were selected as the axioms (basic properties *sine qua non*) for probabilities:

1. Probability of any event is always between 0 and 1: $0 \leqslant \mathbf{P}(A) \leqslant 1$.

2. The certain event Ω has probability 1: $\mathbf{P}(\Omega) = 1$.

3. The probability of the union of mutually exclusive events A_1, A_2, ..., A_n equals the sum of their probabilities:

$$\mathbf{P}(A_1 \cup A_2 \cup \ldots \cup A_n) = \mathbf{P}(A_1) + \mathbf{P}(A_2) + \ldots + \mathbf{P}(A_n). \tag{1.2}$$

The third property can be read informally as “the probability of an event is the total of the probabilities of all its favorable cases”. On the basis of these properties some general laws of probability can be proven.

areas in real and complex analysis, foundations of geometry, functional analysis, and many others.

[5] Sergeĭ Natanovich Bernstein (1880–1968) was a Russian and Soviet mathematician, a Member of USSR Academy of Sciences. He had solved the 19th Hilbert’s problem. His areas of scientific contributions were differential equations, geometry, probability theory, and approximation theory.

[6] Andreĭ Nikolaevich Kolmogorov (1903–1987) was a Soviet mathematician and a Member of USSR Academy of Sciences who made significant contributions to the mathematics of probability theory, topology, intuitionistic logic, turbulence, classical mechanics, algorithmic information theory, and computational complexity.

Table 1.4. Logical operations on random events and their set theory counterparts

Notation	Set theory	Probability theory
$A \cup B$	the union of two sets A, B is a set containing elements which belong either to A or B	a union of two events A, B occurs if either A or B occurs, or both
$A \cap B$	the intersection of two sets A, B is a set containing elements which belong both to A and B	a union of two events A, B occurs if both A and B occur
$\overline{A}$	the complement to the set A	the opposite event which occurs if and only if A didn't occur
Ω	the universe, the set including all other sets as its subsets	the certain event which occurs in every conduction of the experiment
$\varnothing$	the empty set	the impossible event which never occurs
$A \setminus B$	set difference of A and B is a set containing only elements belonging to A and not to B	the difference of events A and B occurs if and only if A occurred without B
$A \subset B$	A is a subset of B	A implies B, A i a favorable case of B, i.e. B occurs whenever A occurs
$A \cap B = \varnothing$	the sets A and B share no common elements	events A and B are mutually exclusive, they can't occur together

Theorem 1. *Probability satisfies the following:*

- *The probability of the impossible event is zero:*
$$\mathbf{P}(\varnothing) = 0.$$
- *The probability of the opposite event:*
$$\mathbf{P}(\overline{A}) = 1 - \mathbf{P}(A). \tag{1.3}$$
- *Monotonicity: if event A implies event B then*
$$\mathbf{P}(A) \leqslant \mathbf{P}(B)$$
and
$$\mathbf{P}(B \setminus A) = \mathbf{P}(B) - \mathbf{P}(A). \tag{1.4}$$
- *Summation formula for two events: if two events A and B are not mutually exclusive then*
$$\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \cap B). \tag{1.5}$$

The summation formula can be proven as follows. The union $A \cup B$ has three mutually exclusive cases: A without B (i.e. $A \setminus B$), B without A (i.e. $B \setminus A$), and both A and B (i.e. $A \cap B$). Then,

$$\mathbf{P}(A \cup B) = \mathbf{P}(A \setminus B) + \mathbf{P}(B \setminus A) + \mathbf{P}(A \cap B).$$

But the intersection $A \cap B$ implies both A and B. Hence,

$$\mathbf{P}(A \setminus B) = \mathbf{P}(A) - \mathbf{P}(A \cap B)$$

and

$$\mathbf{P}(B \setminus A) = \mathbf{P}(B) - \mathbf{P}(A \cap B).$$

Substituting these two equalities info the formula for the union, we get the summation formula:

$$\begin{aligned}
&\mathbf{P}(A \setminus B) + \mathbf{P}(B \setminus A) + \mathbf{P}(A \cap B) \\
&\qquad = \mathbf{P}(A) - \mathbf{P}(A \cap B) + \mathbf{P}(B) - \mathbf{P}(A \cap B) + \mathbf{P}(A \cap B) \\
&\qquad = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \cap B).
\end{aligned}$$

Proofs of the remaining formulae is left to the reader.

Recall that the probability $\mathbf{P}(A)$ of an event A is completely defined by the conditions of the experiment. Let B be an event with positive probability, $\mathbf{P}(B) > 0$.

Definition 1. *The conditional probability of event A given event B with $\mathbf{P}(B) > 0$ is defined as*

$$\mathbf{P}(A \mid B) = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)} . \tag{1.6}$$

If $\mathbf{P}(B) = 0$ then the corresponding conditional probability is undefined. To distinguish, the (unconditional) probability $\mathbf{P}(A)$ is called the absolute probability *of the event A.*

This definition is related to conditional experiments. Assume that the original experiment has been carried out N times. Let us throw away those times when the event B failed to occur. Denote by N_B the number of occurrences of event B. Then in the remaining results we may observe the event A together with B, i.e. we observe only a part of A, namely, $A \cap B$. Denote by $N_{A\cap B}$ the number of occurrences of $A \cap B$. Then what is the meaning of the relative frequency

$$\frac{N_{A\cap B}}{N_B} \, ?$$

It should be the conditional probability of A given B experimentally measured. Then,

$$\frac{N_{A\cap B}}{N_B} = \frac{N_{A\cap B}/N}{N_B/N} \approx \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)} .$$

It justifies the formal definition of conditional probability.

It is left to the reader to prove that conditional probabilities in fact satisfy the axioms on page 13.

From the definition of conditional probabilities we obtain the multiplication theorem for probabilities:

Theorem 2. *Let $\mathbf{P}(A \cap B) > 0$. Then*

$$\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B \mid A) = \mathbf{P}(B)\mathbf{P}(A \mid B). \tag{1.7}$$

Let $\mathbf{P}(A_1 \cap A_2 \cap \ldots \cap A_n) > 0$. Then

$$\begin{aligned}\mathbf{P}(A_1 \cap A_2 \cap \ldots \cap A_n) &= \mathbf{P}(A_1)\mathbf{P}(A_2 \mid A_1)\mathbf{P}(A_3 \mid A_1 \cap A_2) \\ &\quad \times \ldots \times \mathbf{P}(A_n \mid A_1 \cap A_2 \cap \ldots \cap A_{n-1}).\end{aligned} \tag{1.8}$$

Equality (1.7) follows directly from (1.6). To prove (1.8), recursively apply (1.7).

There are special cases when the multiplication theorem involves only absolute probabilities.

Definition 2. *Events A and B are called independent if*

$$\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B). \tag{1.9}$$

Events A_1, A_2, ..., A_n are called independent if

$$\mathbf{P}(A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_k}) = \mathbf{P}(A_{i_1})\mathbf{P}(A_{i_2}) \times \cdots \times \mathbf{P}(A_{i_k}) \tag{1.10}$$

for all $k = 2, 3, \ldots, n$ and $1 \leqslant i_1 < i_2 < \ldots < i_k \leqslant n$. When events are not independent they are called dependent.

Examples demonstrate that independence of two and $n \geqslant 3$ events are not equivalent. For instance, let E_1, E_2, E_3, and E_4 be mutually exclusive events such that $\mathbf{P}(E_1) = \mathbf{P}(E_2) = \mathbf{P}(E_3) = \mathbf{P}(E_4) = 0.25$, set $A = E_1 \cup E_2$, $B = E_1 \cup E_3$, $C = E_1 \cup E_3$. Then A and B are independent, A and C are independent, and B and C are independent, but A, B, and C are dependent:

$$\mathbf{P}(A \cap B \cap C) = \mathbf{P}(E_1) = 0.25 \neq \mathbf{P}(A)\mathbf{P}(B)\mathbf{P}(C) = 0.5 \cdot 0.5 \cdot 0.5.$$

Suppose A and B are independent and $\mathbf{P}(B) > 0$. Then

$$\begin{aligned} \mathbf{P}(A \mid B) &= \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)} && \text{by (1.6)} \\ &= \frac{\mathbf{P}(A)\mathbf{P}(B)}{\mathbf{P}(B)} && \text{by (1.9)} \\ &= \mathbf{P}(A). \end{aligned}$$

So, in probability theory independence means that occurrence of one event doesn't change the probability of the other event. Observation of event B doesn't improve our prediction of chances to observe the other event in the same execution of the experiment. We'd like to stress that the independence in probability theory is a property of probabilities. It is often called a *statistical* independence. Sometimes it is implied by independence in common sense. People usually understand independence as the absence of causal relationships between two phenomena. Statistical independence is more than that.

If A and B are independent, so are A and $\overline{B}$, $\overline{A}$ and B, $\overline{A}$ and $\overline{B}$. Indeed,

$$\begin{aligned} \mathbf{P}(A \cap \overline{B}) &= \mathbf{P}(A \setminus B) \\ &= \mathbf{P}(A \setminus (A \cap B)) \end{aligned}$$

$$
\begin{aligned}
&= \mathbf{P}(A) - \mathbf{P}(A \cap B) && \text{by (1.4)}\\
&= \mathbf{P}(A) - \mathbf{P}(A)\mathbf{P}(B) && \text{by (1.9)}\\
&= \mathbf{P}(A)(1 - \mathbf{P}(B))\\
&= \mathbf{P}(A)\mathbf{P}(\overline{B}) && \text{by (1.3)}
\end{aligned}
$$

If events A_1, A_2, $\ldots$, A_n are independent, so are events A_1', A_2', $\ldots$, A_n' where each A_i' is either A_i or $\overline{A}_i$, $i = 1, 2, \ldots, n$.

When events A_1, A_2, $\ldots$, A_n are independent, the multiplication formula (1.8) takes the following shorter form:

$$\mathbf{P}(A_1 \cap A_2 \cap \ldots \cap A_n) = \mathbf{P}(A_1)\mathbf{P}(A_2) \times \ldots \times \mathbf{P}(A_n). \tag{1.8$'$}$$

Example 3: Probability of at least one of a number of events. Let A_1, A_2, $\ldots$, A_n be independent events, their probabilities be p_1, p_2, $\ldots$, p_n correspondingly. Then at least one of them occurs with probability

$$1 - (1 - p_1)(1 - p_2) \times \ldots \times (1 - p_n).$$

Indeed, the event “at least one of them occurs” can be written as the union $A_1 \cup A_2 \cup \ldots \cup A_n$, so

$$
\begin{aligned}
\mathbf{P}(A_1 \cup A_2 \cup \ldots \cup A_n) &= \mathbf{P}\big(\overline{\overline{A}_1 \cap \overline{A}_2 \cap \ldots \cap \overline{A}_n}\big) && \text{by (1.1)}\\
&= 1 - \mathbf{P}(\overline{A}_1 \cap \overline{A}_2 \cap \ldots \cap \overline{A}_n) && \text{by (1.3)}\\
&= 1 - \mathbf{P}(\overline{A}_1)\mathbf{P}(\overline{A}_2) \times \ldots \times \mathbf{P}(\overline{A}_n) && \text{by (1.8$'$)}\\
&= 1 - (1 - p_1)(1 - p_2) \times \ldots \times (1 - p_n) && \text{by (1.3)}
\end{aligned}
$$

■□■

Combination of the summation axiom 3 and the multiplication theorem gives the *Law of total probability*.

Theorem 3. *Let events H_1, H_2, $\ldots$, H_n be mutually exclusive, $\mathbf{P}(H_i) > 0$ for $i = 1, 2, \ldots, n$, and let $A \subset H_1 \cup H_2 \cup \ldots \cup H_n$, i.e. event A may occur only together with one on the auxiliary events H_1, H_2, $\ldots$, H_n. Then,*

$$\mathbf{P}(A) = \mathbf{P}(H_1)\mathbf{P}(A \mid H_1) + \mathbf{P}(H_2)\mathbf{P}(A \mid H_2) + \ldots + \mathbf{P}(H_n)\mathbf{P}(A \mid H_n). \tag{1.11}$$

Lecture 2 Random variables and their probability distributions

We will not go into details of a purely mathematical definition of a random variable. We assume that the notion of a random variable is one of primary notions in probability theory. It is important to keep in mind that a random variable is described by its set of possible values and probabilities of events generated by it as described in Lecture 1. Let X be a random variable.

Definition 3. *X is a discrete random variable if there are numbers x_1, x_2, …, x_n, … among its possible values that have strictly positive probabilities summing up to unity:*

$$\mathbf{P}(X = x_i) = p_i > 0, \quad i = 1, 2, \ldots, n, \ldots;$$
$$p_1 + p_2 + \ldots + p_n + \ldots = 1.$$

When the possible values are in finite number one may put them in the following table assuming $x_1 < x_2 < \ldots < x_n$:

a	x_1	x_2	$\ldots$	x_n
$\mathbf{P}(X = a)$	p_1	p_2	$\ldots$	p_n

Another visual representation of a discrete probability distribution is the polygon of probabilities (Fig. 2.1). A polygon of probabilities consists of line segments whose end-points are (x_1, p_1), (x_2, p_2), …, (x_n, p_n), … .

Example 4: Discrete uniform probability distribution. Random variable X has a discrete uniform distribution if its values x_1, x_2, …, x_n have equal probabilities of $1/n$. An example of such random variable is the number of points thrown in one roll of a fair die (a symmetric die). One year has 365 days (assuming a common year). We may assume that people are equally likely to be born on any day of year[1]. Hence the day number X has a uniform probability distribution on integers 1 through 365:

$$\mathbf{P}(X = k) = \frac{1}{365}, \qquad k = 1, 2, \ldots, 365.$$

[1] Some authors argue that September is more frequent than the other months in the USA, see http://www.panix.com/~murphy/bday.html

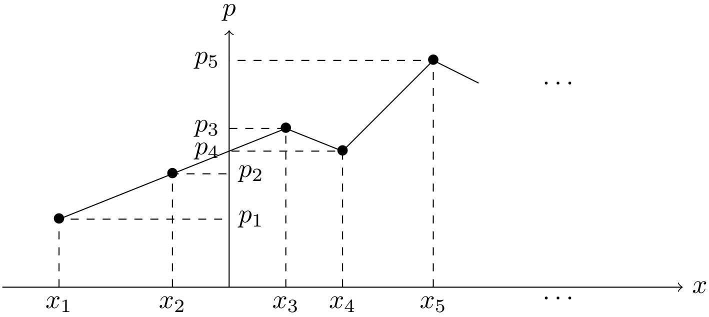

Figure 2.1. A polygon of probabilities. Only the first five points are depicted

Then the probability to born in January is by (1.2)

$$\mathbf{P}(1 \leqslant X \leqslant 31) = \mathbf{P}(X = 1) + \mathbf{P}(X = 2) + \ldots + \mathbf{P}(X = 31) = \frac{31}{365}\,.$$

It is timely to discuss the rare events comprehension. In short, we can't neglect events of small probability. Consider a well-shuffled deck of 52 cards. If we pick 13 cards at random they can be any cards. In other words, no combination of 13 cards is more likely to appear than the others. Let us enumerate all possible combinations (they are

$$\binom{52}{13} = \frac{52!}{13!(52-13)!} = 635\,013\,559\,600,$$

i.e. more than 635 billion) and let the random variable X indicate the number of the combination at hand. Then X has the uniform probability distribution on the set of integers 1, 2, $\ldots$, 635 013 559 600. Thus every combination a of thirteen cards has probability as small as

$$\mathbf{P}(X = a) = \frac{1}{635\,013\,559\,600} \approx 1.6 \times 10^{-12}.$$

So, every single value of this random variable is extremely rare. But we can't get rid of these values because there are no other "more likely" values, and what is more important, such rare events are favorable cases for more likely events. For instance, there are

$$\binom{48}{13} = 192\,928\,249\,296$$

combinations of thirteen cards without Aces, so the probability to have no Aces in the hand is

$$\begin{aligned}\mathbf{P}(\{a\colon \text{the combination } a \text{ has no Aces}\}) &= \frac{192\,928\,249\,296}{635\,013\,559\,600} \\ &= 0.3038175270108043\ldots\end{aligned}$$

and is expected in around 30 % of experiments. ■□■

Example 5: Binomial probability distribution. Random variable X with possible values 0, 1, …, n and corresponding probabilities

$$\begin{aligned}\mathbf{P}(X = k) &= \binom{n}{l} p^k (1-p)^{n-k} \\ &= \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}, \qquad k = 0, 1, \ldots, n \qquad (2.1)\end{aligned}$$

has a binomial probability distribution. Binomial probability distribution describes for instance the number of occurrences of an event A with probability p in n repeated independent experiments.

Assume a symmetric coin is tossed n times. Select a face, its appearance will be called a success. The symmetry implies the probability of a success in one toss $p = 0.5$. A coin is a solid body, neither its form or its mass distribution changes from a toss to another, hence the tosses are independent, they are executed in the same controlled conditions. Then the number of occurrences of the selected face follows the binomial probability distribution. Substituting n and $p = 0.5$ into formula (2.1) we get:

$$\begin{aligned}\mathbf{P}(X = 2) &= \binom{4}{2} \times (0.5)^4 = 0.375 \qquad \text{if } n = 4, \\ \mathbf{P}(X = 5) &= \binom{10}{5} \times (0.5)^{10} = 0.24609375 \qquad \text{if } n = 10, \\ \mathbf{P}(X = 25) &= \binom{50}{25} \times (0.5)^{50} = 0.11227517\ldots \qquad \text{if } n = 50, \\ \mathbf{P}(X = 50) &= \binom{100}{50} \times (0.5)^{100} = 0.07958923\ldots \qquad \text{if } n = 100, \\ \mathbf{P}(X = 500) &= \binom{1000}{500} \times (0.5)^{1000} = 0.025225018\ldots \qquad \text{if } n = 1000.\end{aligned}$$

We see that the probability of even divide between Heads and Tails becomes less and less probable as the number of tosses grows. ■□■

Example 6: Geometric probability distribution.. Random variable X with possible values 1, 2, ... and probabilities

$$\mathbf{P}(X = k) = p(1-p)^{k-1}, \qquad k = 1, 2, \dots$$

has the geometric probability distribution with parameter p, $0 < p \leqslant 1$ (if $p = 1$ then $\mathbf{P}(X = 1) = p$ by definition). Such random variables arise as the count of independent repeated trials until the first occurrence of a selected event A, p being the probability of A in one trial. E.g., consider customers arrivals in discrete time-scale. If one customer arrives during a time slot with probability p and no customers arrive with probability $1 - p$, then X is the interarrival interval.

The following computation demonstrates the lack of memory property of the geometric probability distribution. Let k and l be positive integers, then one has by (1.2)

$$\begin{aligned}\mathbf{P}(X > k) &= \mathbf{P}(X = k+1) + \mathbf{P}(X = k+2) + \mathbf{P}(X = k+3) + \dots \\ &= p(1-p)^k + p(1-p)^{k+1} + p(1-p)^{k+2} + \dots \\ &= \frac{p(1-p)^k}{1-(1-p)} = (1-p)^k \end{aligned} \tag{2.2}$$

by the sum formula for a geometric progression; similarly,

$$\mathbf{P}(X > k + l) = (1-p)^{k+l}.$$

Now,

$$\begin{aligned}\mathbf{P}(X > k+l | X > k) &= \frac{\mathbf{P}(\{X > k+l\} \cap \{X < k\})}{\mathbf{P}(X > k)} && \text{by (1.6)} \\ &= \frac{\mathbf{P}(\{k+l+1, k+l+2, \dots\} \cap \{k+1, k+2, \dots\})}{\mathbf{P}(\{k, k+1, \dots\})} \\ &= \frac{\mathbf{P}(\{k+l+1, k+l+2, \dots\} \cap \{k+1, k+2, \dots\})}{\mathbf{P}(\{k, k+1, \dots\})} \\ &= \frac{\mathbf{P}(\{k+l+1, k+l+2, \dots\})}{\mathbf{P}(\{k, k+1, \dots\})} \\ &= \frac{(1-p)^{k+l}}{(1-p)^k} && \text{by (2.2)} \\ &= (1-p)^l = \mathbf{P}(X > l). \end{aligned}$$

The conditional probability above refers to the remaining waiting time being at least $l+1$ given no successes occurred during the first k trials. By the way, equality (2.2) can be discovered as follows. Since the event $X > k$ means that the first k trials result in a failure and the trials are independent (2.2) follows from the multiplication formula (1.8′) for probabilities of independent events, $1-p$ being the probability of a failure in one trial.
■□■

Example 7: Poisson probability distribution. A random variable X is said to have the Poisson[2] probability distribution with parameter λ, $\lambda > 0$ if it takes on values $k = 0, 1, 2, \ldots$ with probabilities

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

This probability distribution is widely used in insurance mathematics as well as in quality control. It well describes the number of rare events in a large number of independent trials. The Poisson probability distribution is a good approximation for the binomial probability distribution. Indeed, let n grow to infinity and p tend to 0 so that $np \to \lambda$. Put

$$R(k; n, p) = \frac{\binom{n}{k} p^k (1-p)^{n-k}}{\binom{n}{k-1} p^{k-1} (1-p)^{n-(k-1)}} = \frac{1-p}{p} \times \frac{n-k+1}{k},$$

then

$$\begin{gathered}\binom{n}{k} p^k (1-p) n - k = (1-p)^n R(1; n, p) R(2; n, p) \cdots R(k; n, p), \\ R(k; n, p) \to \frac{\lambda}{k} \qquad \text{as } n \to \infty,\ p \to 0, \text{ and } np \to \lambda, \\ (1-p)^n \to e^{-\lambda},\end{gathered}$$

and

$$\binom{n}{k} p^k (1-p)^{n-k} \to \frac{\lambda^k}{k!} e^{-\lambda}.$$

The latter limit has the name of Poisson's limit theorem.

[2] Simeón Denis Poisson [French: pwa.sɔ̃] (1781–1840) was a French mathematician and physicist. His works cover areas of pure mathematics, mathematical physics, theoretical and celestial mechanics.

Consider the following case study: a refrigerator production lines produces on the average defective refrigerators per ten thousand. What is the probability a party of 500 refrigerators contains at least one defective refrigerator? From the data we may estimate the probability to produce a defective refrigerator as $3/10\,000 = 0.0003$. The number of trials is $n = 500$, a defective refrigerator will be counted as a 'success', then the number X of defective refrigerators in the party has the Poisson probability distribution with parameter $\lambda = 500 \times 0.0003 = 0.15$. So,

$$\mathbf{P}(X \geqslant 1) = \sum_{k=1}^{\infty} \frac{0.15^k}{k!} e^{-0.15}.$$

This series is not easy to evaluate directly. But, turning to the opposite event one obtains

$$\mathbf{P}(X \geqslant 1) = 1 - \mathbf{P}(X = 0) = 1 - e^{-0.15} = 0.1392920235749422\ldots$$

In other words, on the average only 86.1 % of such parties are completely high-quality. It looks like a paradox, since the probability to buy a defective refrigerator is still 0.0003, i.e. is 'negligibly small'. ■□■

The second important type of random variables is that of continuous random variables.

Definition 4. *A random variable X with an uncountable infinite set of possible values is called a continuous random variable if for any u, v $(u < v)$ the probability $\mathbf{P}(u < X < v)$ can be evaluated as*

$$\mathbf{P}(u < X < v) = \int\limits_{u}^{v} p(x)\, dx, \tag{2.3}$$

where $p(u)$ is a non-negative function such that

$$\int\limits_{-\infty}^{\infty} p(u)\, du = 1. \tag{2.4}$$

The function $p(u)$ is called a probability density. It can be determined by taking limit

$$p(u) = \lim_{\substack{\delta \to 0, \delta > 0 \\ \eta \to 0, \eta > 0}} \frac{\mathbf{P}(u - \delta < X < u + \eta)}{\eta + \delta}\,. \tag{2.5}$$

Its interpretation is as follows:

$$\mathbf{P}(u-\delta < X < u+\eta) = p(u)(\delta+\eta) + \dots$$

where the ellipsis stands for the small terms of higher order then the total length $\eta+\delta$, i.e. the probability for the random variable X to take a value between the limits $u-\delta$ and $u+\eta$ is approximately the area of the rectangle with sides $p(u)$ and $(\delta+\eta)$ (see Fig. 2.2).

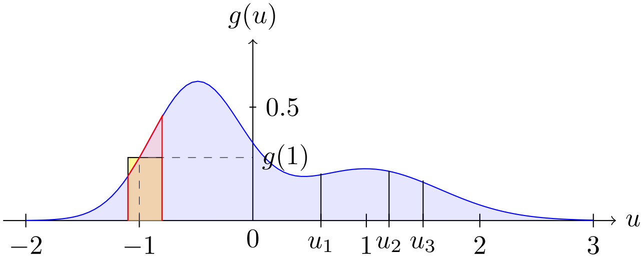

Figure 2.2. Probability density and geometric visualization of additivity property of probabilities. The blue curve depicts the probability density. The rectangle's area is approximately equal to the area below the curve within the same limits

Further, the general properties of definite integrals guarantee that for $u_1 < u_2 < u_3$

$$\mathbf{P}(u_1 < X < u_3) = \mathbf{P}(u_1 < X < u_2) + \mathbf{P}(u_2 < X < u_3)$$

where

$$\mathbf{P}(u_1 < X < u_3) = \int\limits_{u_1}^{u_3} p(u)\,du,$$

$$\mathbf{P}(u_1 < X < u_2) = \int\limits_{u_1}^{u_2} p(u)\,du,$$

$$\mathbf{P}(u_2 < X < u_3) = \int\limits_{u_2}^{u_3} p(u)\,du.$$

Let us calculate the probability of any single value of a continuous random variable. Call this value u. Then the event $X = u$ implies events

$u - \delta < X < u + \eta$ for all strictly positive δ and η. Hence, by the monotonicity of probability (see Lecture 1),

$$0 \leqslant \mathbf{P}(X = u) \leqslant \mathbf{P}(u - \delta < X < u + \eta) = \int\limits_{u-\delta}^{u+\eta} p(u)\, du.$$

But the definite integral in the right-hand side can be made arbitrary small by choosing small values of δ and η. We must conclude that $\mathbf{P}(X = u) = 0$. Again, we have an abundance of events of zero probability none of which can be got rid of. Another implication of the conclusion is that strict inequalities and weak inequalities have same probabilities, e.g.

$$\mathbf{P}(X \leqslant u) = \mathbf{P}(X < u) + \mathbf{P}(X = u) = \mathbf{P}(X < u),$$
$$\mathbf{P}(u \leqslant X \leqslant v) = \mathbf{P}(u < X \leqslant v) = \mathbf{P}(u \leqslant X < v) = \mathbf{P}(u < X < v).$$

Example 8: Continuous uniform probability distribution. Let the density $p(u)$ take on a positive value inside an interval (a, b) and vanish outside of it, i.e.

$$p(u) = \begin{cases} C & \text{if } a \leqslant u \leqslant b \\ 0 & \text{if } u < a \text{ or } u > b. \end{cases}$$

To determine the constant C we may use the normalization condition (2.4):

$$\begin{aligned} 1 &= \int\limits_{-\infty}^{\infty} p(u)\, du \\ &= \int\limits_{-\infty}^{a} p(u)\, du + \int\limits_{a}^{b} p(u)\, du + \int\limits_{b}^{\infty} p(u)\, du \\ &= 0 + \int\limits_{a}^{b} p(u)\, du + 0 = C \cdot (b - a), \end{aligned}$$

whence $C = 1/(b - a)$. This probability distribution as analogous to its discrete counterpart. It realizes the idea of equal probabilities of cases. Although each single value has zero probability, every segment of length Δ inside the interval (a, b) has the same probability:

$$\mathbf{P}(u < X < u + \Delta) = \int\limits_{u}^{u+\Delta} \frac{du}{b - a}$$

$$= \frac{u}{b-a}\bigg|_{u}^{u+\Delta} = \frac{\Delta}{b-a} \tag{2.6}$$

In other words, the probability of hitting an interval or a segment is proportional only to the length of that interval or segment. ■□■

Example 9: Exponential probability distribution. The exponential probability distribution with parameter λ, $\lambda > 0$ is defined by the probability density

$$p(u) = \begin{cases} 0 & \text{if } u < 0, \\ \lambda e^{-\lambda u} & \text{if } u \geqslant 0. \end{cases}$$

■□■

Example 10: Pareto probability distribution. The Pareto probability distribution is defined by the density

$$p(u) = \begin{cases} 0 & \text{if } u < 0, \\ \dfrac{\alpha\theta^{\alpha}}{(u+\theta)^{\alpha+1}} & \text{if } u \geqslant 0. \end{cases}$$

The parameters are $\alpha > 0$ and $\theta > 0$. This probability distribution is often used to model the distribution of wealth in a population. It is instructive to compare a Pareto probability density to an exponential probability density (see Fig. 2.3). Although both types of probability densities are monotonously decreasing for $u > 0$, exponential probability densities decay more rapidly. Informally, mathematicians say that Pareto densities are 'heavy-tailed'.

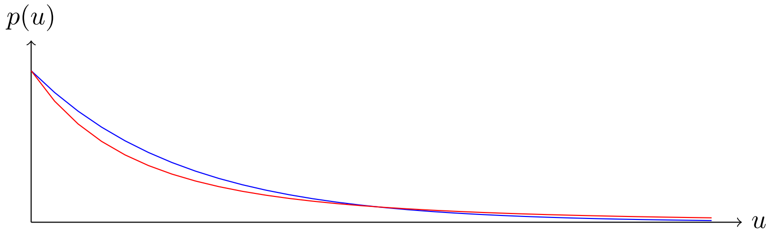

Figure 2.3. A Pareto probability density (the red curve) and exponential probability density (the blue curve)

■□■

Example 11: Normal (Gaussian) probability distribution. The normal probability distribution (also known as the Gaussian probability distribution) has the probability density

$$p(u) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(u-a)^2}{2\sigma^2}}\,, \tag{2.7}$$

where $\sigma > 0$ and a real a are the parameters. This probability distribution will be more discussed in Lecture 5. ■□■

Definition 5. *The mathematical expectation of a discrete random variable X is defined as*

$$\mathbf{M}X = x_1p_1 + x_2p_2 + \ldots + x_np_n$$

when the total number of possible value is finite, and the infinite series

$$\mathbf{M}X = x_1p_1 + x_2p_2 + \ldots + x_np_n + \ldots$$

when the total number of possible values in infinite and the series is absolutely convergent, which means the series

$$|x_1|\, p_1 + |x_2|\, p_2 + \ldots + |x_n|\, p_n + \ldots < \infty.$$

If $y = g(x)$ is a function, then

$$\mathbf{M}g(X) = g(x_1)p_1 + g(x_2)p_2 + \ldots + g(x_n)p_n$$

in case of a finite set of possible values, otherwise

$$\mathbf{M}g(X) = g(x_1)p_1 + g(x_2)p_2 + \ldots + g(x_n)p_n + \ldots$$

assuming the series in the right-hand side is absolutely convergent. Let $\mathbf{M}X = a$. Then

$$\mathbf{var}\, X = \mathbf{M}(X-a)^2 = (x_1-a)^2p_1 + (x_2-a)^2p_2 + \ldots$$

is called the variance of X.

The idea to introduce the mathematical expectation comes from the following reasoning. Let numbers x_1, x_2, …, x_n be the possible values of a discrete random variable X, $p_k = \mathbf{P}(X = x_k)$ for $k = 1$, 2, …, n. Denote by N_k the number of occurrences of x_k in N independent trials. Then the *sample average value* of X is

$$x_1 \cdot \frac{N_1}{N} + x_2 \cdot \frac{N_2}{N} + \ldots + x_n \cdot \frac{N_n}{N}$$

On the other hand, the statistical stability leads to

$$\frac{N_k}{N} \approx p_k.$$

So, sample average value should be close to

$$\mathbf{M}X = x_1 p_1 + x_2 p_2 + \ldots + x_n p_n.$$

Definition 6. *The mathematical expectation of a continuous random variable X with probability density $p(u)$ is defined as*

$$\mathbf{M}X = \int\limits_{-\infty}^{\infty} u\, p(u)\, du$$

the integral is absolutely convergent, which means

$$\int\limits_{-\infty}^{\infty} |u|\, p(u)\, du < \infty.$$

If $y = g(x)$ is a function, then

$$\mathbf{M}g(X) = \int\limits_{-\infty}^{\infty} g(u) p(u)\, du$$

assuming the integral in the right-hand side is absolutely convergent. Let $\mathbf{M}X = a$. Then the variance of X is defined as

$$\mathbf{var}\, X = \mathbf{M}(X - a)^2 = \int\limits_{-\infty}^{\infty} (u - a)^2 p(u)\, du.$$

Mathematical expectations of discrete and continuous random variables have different definitions. But then one defines the variance by same formula regardless of the nature of the random variable. This is a general situation when more advanced quantities are defined in terms of mathematical expectation.

Definition 7. *The standard deviation of a random variable X is defined as*

$$\sigma(X) = \sqrt{\mathbf{var}\, X}.$$

Just as the mathematical expectation locates the random variable and characterizes its mean value, the variance demonstrates typical deviation of the value taken on by the random variable from its mean value. Both mathematical expectation and variance play important role in theory and applications.

Example 12: Measuring the risk of an asset. Different assets such as shares, security papers, options, and futures are sold and bought at financial markets and stock exchanges worldwide. Their prices change randomly every moment. Let S_t be the closing price of a selected asset at day $t = 1, 2, \ldots$. The *return* (or the *interest rate*) at day t is defined as

$$R_t = \frac{S_t - S_{t-1}}{S_{t-1}} \times 100\ \%.$$

People are often interested in *expected return* $\mathbf{M}R_t$ of the asset. Its risk is defined as $\sigma(R_t)$, its stand-art deviation. These data are published in financial sections of newspapers and TV news channels etc. Let the return have the following discrete probability distribution:

x	9 %	12 %	15 %
$\mathbf{P}(R_t = x)$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

Then

$$\begin{aligned}
\mathbf{M}R_t &= \Big(0.09 \times \frac{1}{3} + 0.12 \times \frac{1}{3} + 0.15 \times \frac{1}{3}\Big) \times 100\ \% = 12\ \%,\\
\mathbf{var}\, R_t &= (0.09 - 0.12)^2 \times \frac{1}{3} + (0.12 - 0.12)^2 \times \frac{1}{3} + (0.15 - 0.12)^2 \times \frac{1}{3}\\
&= 0.0006,\\
\sigma(R_t) &= \sqrt{0.0006} \times 100\ \% = 2.45\ \%.
\end{aligned}$$

■□■

In computation of mathematical expectations and variances more formulae are helpful.

Theorem 4. *Let X and Y be random variables with finite* $\mathbf{M}X$ *and* $\mathbf{M}Y$*, let a and b be constants, then*

$$\mathbf{M}a = a, \quad \mathbf{var}\,(a) = 0, \tag{2.8}$$

$$\mathbf{M}(aX + bY) = a\mathbf{M}X + b\mathbf{M}Y, \tag{2.9}$$

$$\mathbf{var}\, X = \mathbf{M}(X^2) - (\mathbf{M}X)^2, \tag{2.10}$$

$$\mathbf{var}\, X = \mathbf{M}X(X-1) + \mathbf{M}X - (\mathbf{M}X)^2, \tag{2.11}$$

$$\mathbf{var}\,(aX+b) = a^2\mathbf{var}\,X, \tag{2.12}$$

$$\mathbf{var}\,(X+Y) = \mathbf{var}\,X + \mathbf{var}\,Y + 2\mathbf{M}\big((X-\mathbf{M}X)(Y-\mathbf{M}Y)\big). \tag{2.13}$$

Recall our discussion of independence in the previous lecture. Now we may ask a question: when probabilities

$$\mathbf{P}(u_1 < X < u_2) \quad \text{and} \quad \mathbf{P}(v_1 < Y < v_2)$$

are sufficient for calculation of the probability

$$\mathbf{P}(\{u_1 < X < u_2\} \cap \{v_1 < Y < v_2\}) = \mathbf{P}(u_1 < X < u_2, v_1 < Y < v_2)\,?$$

In other words, when are events $u_1 < X < u_2$ and $v_1 < Y < v_2$ are independent? Independence may take place only for specific values u_1, u_2, v_1, and v_2, or it may hold for all $u_1 < u_2$ and $v_1 < v_2$. Geometrically, inequalities

$$u_1 < X < u_2 \qquad v_1 < Y < v_2$$

describe a rectangle with sides parallel to the coordinate axes. If one can easily compute probability for rectangles, one can compute probabilities for many other plain figures which can be approximated by rectangles (see Fig. 2.4).

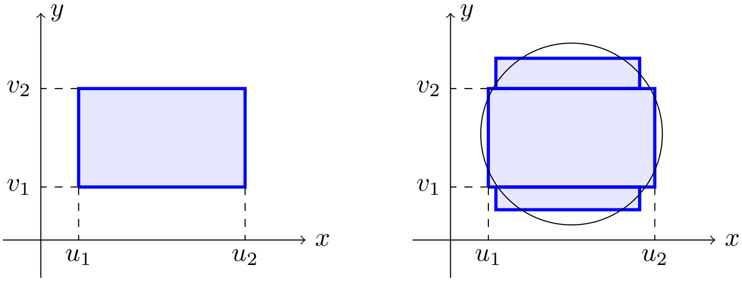

Figure 2.4. Geometrical interpretation of inequalities $u_1 < X < u_2$ and $v_1 < Y < v_2$, and covering plain figures with rectangles

Definition 8. *Random variables X, Y, …, Z are called independent when for all $u_1 < u_2$, $v_1 < v_2$, …, $w_1 < w_2$ one has*

$$\begin{aligned}\mathbf{P}(u_1 < X < u_2, v_1 < Y < v_2, \ldots, w_1 < Z < w_2)\\ = \mathbf{P}(u_1 < X < u_2)\mathbf{P}(v_1 < Y < v_2) \times \ldots \times \mathbf{P}(w_1 < Z < w_2).\end{aligned}$$

Let X and Y be independent continuous random variables with probability densities $p_1(u)$ and $p_2(u)$. Then

$$\mathbf{P}(u_1 < X < u_2, v_1 < Y < v_2) = \mathbf{P}(u_1 < X < u_2)\mathbf{P}(v_1 < Y < v_2)$$
$$= \Big(\int_{u_1}^{u_2} p_1(u)\, du\Big)\Big(\int_{v_1}^{v_2} p_2(v)\, dv\Big) \iint\limits_{\substack{u_1<u<u_2\\ v_1<v<v_2}} p_1(u)p_2(v)\, dudv.$$

Therefore, the probability to hit the rectangle with a random point (X, Y) can be calculated by taking the double integral over the rectangle of the function

$$p(u, v) = p_1(u)p_2(v)$$

of two real variables. In multivariate calculus it is proven that for large class of plain figures D the probability of hitting a figure D equals the double integral

$$\iint\limits_{(u,v)\in D} p(u, v)\, dudv.$$

Definition 9. *Random variables X, Y, ..., Z with uncountable infinite sets of possible values have a continuous joint probability distribution if for any solid figure D the probability*

$$\mathbf{P}((X, Y, \dots, Z) \in D) = \int\limits_{D}\!\cdots\!\int p(u, v, \dots, w)\, du\, dv \cdots dw \tag{2.14}$$

where $p(u, v, \dots, w)$ is a non-negative function such that

$$\int\limits_{\substack{-\infty<u<\infty\\ -\infty<v<\infty\\ \dots\\ -\infty<w<\infty}}\!\cdots\!\int p(u, v, \dots, w)\, du\, dv \cdots dw = 1.$$

The function $p(u, v, \dots, w)$ is called a multivariate probability density (or a joint probability density).

Random variables X, Y, ..., Z with joint continuous probability distribution are independent if and only if

$$p(u, v, \dots, w) = p_1(u)p_2(v) \times \cdots \times p_n(w)$$

where $p_1(u)$ is the probability density of X, $p_2(v)$ is the probability density of Y, $p_n(w)$ is the probability density of Z.

Equivalently, discrete random variables X, Y, …, Z are independent if and only if

$$\mathbf{P}(X = u, Y = v, \ldots, Z = w) = \mathbf{P}(X = u)\mathbf{P}(Y = v) \times \ldots \times \mathbf{P}(Z = w)$$

for all possible values u, v, …, w of the random variables.

One shouldn't think that random variables are usually independent.

Example 13: Random sampling without replacement. From a set or n items k items are selected randomly[3]. Assuming the items were labelled 1 to n and the selected items were ordered by their labels, denote by X_1, X_2, …, X_k the labels from the smallest to the largest. Obviously, $X_1 < X_2 < \ldots < X_k$. Besides that, all probabilities

$$\mathbf{P}(X_1 = u_1, X_2 = u_2, \ldots, X_k = u_k), \quad 1 \leqslant u_1 < u_2 < \ldots < u_k \leqslant n$$

should be equal. Since there are $\binom{n}{k}$ different sequences $(u_1, u_2, \ldots, u_k)$, $1 \leqslant u_1 < u_2 < \ldots < u_k \leqslant n$ and the sum of all such probabilities equals unity as the probability of the certain event, we get

$$\mathbf{P}(X_1 = u_1, X_2 = u_2, \ldots, X_k = u_k) = \begin{cases} \dfrac{1}{\binom{n}{k}} & \text{if } 1 \leqslant u_1 < u_2 < \ldots < u_k \leqslant n \\ 0 & \text{otherwise.} \end{cases}$$

In particular,

$$\mathbf{P}(X_1 = n - k + 1, X_2 = n - k + 2, \ldots, X_k = n) = \frac{k!(n-k)!}{n!} > 0.$$

The event $X_1 = n-k+1, X_2 = n-k+2, \ldots, X_k = n$ implies $X_1 = n-k+1$ and

$$\mathbf{P}(X_1 = n - k + 1) \geqslant \mathbf{P}(X_1 = n - k + 1, X_2 = n - k + 2, \ldots, X_k = n) > 0.$$

Then

$$\mathbf{P}(X_1 = n - k, X_2 = n - k + 1, \ldots, X_k = n - 1) = \frac{k!(n-k)!}{n!} > 0.$$

[3] To choose an item at random one can use a roulette with sufficiently many pockets. If a pocket appears more than once, it's skipped for another turn of the roulette.

The event $X_1 = n-k, X_2 = n-k, \ldots, X_k = n-1$ implies $X_2 = n-k+1$ and

$$\mathbf{P}(X_2 = n-k+1) \geqslant \mathbf{P}(X_1 = n-k, X_2 = n-k+1, \ldots, X_k = n-1) > 0.$$

This way we notice that all $\mathbf{P}(X_i = n-k+1) > 0$, $i = 1, 2, \ldots, k$. But

$$\begin{aligned} 0 &= \mathbf{P}(X_1 = n-k+1, X_2 = n-k+1, \ldots X_k = n-k+1) \\ &\neq \mathbf{P}(X_1 = n-k+1)\mathbf{P}(X_2 = n-k+1) \\ &\quad \times \ldots \times \mathbf{P}(X_k = n-k+1) > 0. \end{aligned}$$

It means that the random variables X_1, X_2, $\ldots$, X_k are dependent. In this particular example dependence could be seen also from the fact the range of possible values of X_2 is $\{X_1+1, X_1+2, \ldots, n-k+2\}$ and depends on the value taken on by X_1. Random sampling without replacement is important for quality control and the example will be continued in next lectures. ■□■

Independent variables have special properties with respect to their mathematical expectation and variances.

Theorem 5. *If random variables X and Y are independent then*

$$\mathbf{M}(XY) = (\mathbf{M}X)(\mathbf{M}Y), \tag{2.15}$$

$$\mathbf{var}\,(X+Y) = \mathbf{var}\,X + \mathbf{var}\,Y. \tag{2.16}$$

Random variables don't necessarily need to be independent for (2.15) to hold.

Definition 10. *Random variables X and Y are called uncorrelated if*

$$\mathbf{M}(XY) = (\mathbf{M}X)(\mathbf{M}Y).$$

To compute the mathematical expectation of the product XY by definition the probability distribution of the product should be known. The problem of finding it will be discussed in the next lecture. It turns out that the mathematical expectation can also be calculated from the joint probability distribution of X and Y.

Theorem 6. *Let $g(u, v, \ldots w)$ be a real-valued function of real variables u, v, $\ldots$, w. If random variables X, Y, $\ldots$, Z are discrete then*

$$\begin{aligned} \mathbf{M}g(X, Y, \ldots, Z) = \sum_{(u,v,\ldots,w)} & g(u, v, \ldots, w) \\ & \times \mathbf{P}(X = u, Y = v, \ldots, Z = w). \end{aligned}$$

If random variables X, Y, ..., Z are continuous with joint probability density $p(u, v, \dots, w)$ then

$$\mathbf{M}g(X, Y, \dots, Z) = \int \cdots \int_{\substack{-\infty<u<\infty \\ -\infty<v<\infty \\ \cdots \\ -\infty<w<\infty}} g(u, v, \dots, w)p(u, v, \dots, w)\, dudv \cdots dw.$$

Let $\mathbf{M}X = a$, $\mathbf{M}Y = b$ then the quantity

$$\mathbf{cov}\,(X, Y) = \mathbf{M}\big((X - a)(Y - b)\big)$$

is called the covariance between X and Y.

Example 14: Planar Gaussian distribution. Random variables X, Y with joint probability density

$$p(u, v) = \frac{1}{\sigma_1\sigma_2 2\pi\sqrt{1 - r^2}} e^{-f(u,v)/2(1-r^2)}, \tag{2.17}$$
$$f(u, v) = \Big(\frac{u - a}{\sigma_1}\Big)^2 - 2r\frac{(u - a)(v - b)}{\sigma_1\sigma_2} + \Big(\frac{v - b}{\sigma_2}\Big)^2$$

are said to have the two-dimensional normal (Gaussian) probability distribution. Then

$$\mathbf{M}X = a, \quad \mathbf{M}Y = b, \quad \mathbf{var}\,X = \sigma_1^2, \quad \mathbf{var}\,Y = \sigma_2^2, \quad \mathbf{cov}\,(X, Y) = r\sigma_1\sigma_2.$$

The random variables are independent if and only if $r = 0$. ■□■

Example 15: Counterexample to the above. Let

$$p_1(u, v) = \frac{e^{-(u^2-2r_1uv+v^2)/2(1-r_1^2)}}{2\pi\sqrt{1 - r_1^2}}, \quad p_2(u, v) = \frac{e^{-(u^2-2r_2uv+v^2)/2(1-r_2^2)}}{2\pi\sqrt{1 - r_2^2}}$$

two normal densities in two dimensions. Then

i) $\frac{1}{2}(p_1(u, v) + p_2(u, v))$ is a joint probability density;

ii) the joint probability distribution is not a planar normal normal probability density;

iii) each of the random variables with this joint probability density has a normal distribution and they are dependent. ■□■

Table 2.1: Important probability distributions and their characteristics

Name	probability density, values	parametric set	mathematical expectation	variance
uniform discrete	$\frac{1}{n}$ $k = 1,\, 2,\, \ldots,\, n$	$n = 1,\, 2,\, \ldots$	$\frac{n+1}{2}$	$\frac{n^2-1}{2}$
binomial	$\binom{n}{k} p^k (1-p)^{n-k}$ $k = 0,\, 1,\, \ldots,\, n$	$n = 1,\, 2,\, \ldots$ $0 < p < 1$	np	$np(1-p)$
geometric	$p(1-p)^{k-1}$ $k = 0,\, 1,\, \ldots$	$0 < p \leqslant 1$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Poisson	$\frac{\lambda^k}{k!} e^{-\lambda}$ $k = 0,\, 1,\, \ldots$	$\lambda > 0$	λ	λ
hypergeometric	$\frac{\binom{m}{j}\binom{n-m}{k-j}}{\binom{n}{k}}$ $j = 0,\, 1,$ $\ldots,\, \min\{k, m\}$	$1 \leqslant m \leqslant n$ $1 \leqslant k \leqslant n$	$\frac{km}{n}$	$\frac{m(n-m)}{n^2}$ $\times \frac{k(n-k)}{n-1}$
negative binomial	$\binom{k+r-1}{k} p^r (1-p)^k$ $k = 0,\, 1,\, \ldots$	$0 < p \leqslant 1$ $r = 1,\, 2,\, \ldots$	$\frac{(1-p)r}{p}$	$\frac{(1-p)r}{p^2}$
uniform continuous	$p(u) = \frac{1}{b-a}$ $a \leqslant u \leqslant b$	$-\infty < a$ $< b < \infty$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
exponential	$p(u) = \lambda e^{-\lambda u}$ $u > 0$	$\lambda > 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Gaussian (normal)	$\frac{1}{\sigma\sqrt{2\pi}} e^{-(u-a)^2/2\sigma^2}$ $-\infty < u < \infty$	$-\infty < a < \infty$ $\sigma > 0$	a	σ^2

Continued on next page

Name	probability density, values	parametric set	mathematical expectation	variance
χ^2_n, chi-square	$\frac{u^{n/2-1}e^{-u/2}}{2^{n/2}\Gamma(n/2)}$ $u > 0$	$n = 1,\ 2,\ \dots$	n	$2n$
Cauchy	$\frac{1}{\pi(1+(u-a)^2)}$ $-\infty < u < \infty$	$-\infty < a < \infty$	—	—
Student's t	$\frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi n}\,\Gamma\left(\frac{n}{2}\right)}$ $\times\left(1+\frac{u^2}{n}\right)^{-\frac{n+1}{2}}$ $n > 0$	$-\infty < u < \infty$	n if $n > 1$ —, $n \leqslant 1$	$\frac{n}{n-2}$, $n > 2$ ∞ if $n = 2$ —, $n < 2$
Fisher's F	$\left(\frac{m}{n}\right)^{m/2}\frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)}$ $\times\frac{u^{m/2-1}}{\left(1+\frac{mu}{n}\right)^{(m+n)/2}}$ $m, n > 0$	$u > 0$	$\frac{n}{n-2}$, $n > 2$	$\frac{2n^2}{m(m-2)^2}$ $\times\frac{m+n-2}{n-4}$ $n > 4$

Lecture 3 | Probability distributions of functions of several random variables

When constructing mathematical models for real-world phenomena (i.e. in the field of Economics) one searches first for functional relations between variables. One seeks how the 'output' variables of the model depend on its 'input' variables. A typical problem here is to determine the law of probability distribution of the output variables given the laws of probability distribution of the input variables.

Example 16: Hypergeometric distribution, continuation of Example 13. A milk-bottling machine produced n bottles and k of them are selected at random for quality control. Among n there are m flawed bottle. Let X be the number of flawed bottle in the sample for the quality control. What is the probability distribution of X? The settings for the experiment are exactly those from Example 13. Assume additionally that flawed bottles are 1 through m. Then $X_i = j$ for $j \leqslant m$ means that the i-th bottle is flawed while the same event for $j > m$ means the i-th bottle is good. The random variables X_1, X_2, …, X_k are input variables. Then X is an output variable defined as follows:

$$\begin{aligned} &X = 0 && \text{if } X_1 > m;\\ &X = j && \text{if } X_j \leqslant m \text{ and } X_{j+1} > m,\ j = 1, 2, \ldots, k-1;\\ &X = k && \text{if } X_k \leqslant m. \end{aligned}$$

It is a discrete random variable with values 0, 1, …, k. To evaluate probabilities $\mathbf{P}(X = j)$ for $j = 0$, 1, …, k we have to sum according to formula (1.2) probabilities of all favorable cases, i.e. events

$$\{X_1 = u_1, X_2 = u_2, \ldots, X_k = u_k\}$$

such that

$$1 \leqslant u_1 < u_2 < \ldots < u_j \leqslant m < u_{j+1} < \ldots < u_k \leqslant n.$$

According to the rules of combinatorics, there are exactly $\binom{m}{j}$ combinations of u_1, u_2, …, u_j and $\binom{n-m}{k-j}$ combinations of u_{j+1}, u_{j+2}, …, u_k satisfying these conditions. Since each favorable case has probability $\binom{n}{k}^{-1}$, the sum giving the desired probability is

$$\mathbf{P}(X = j) = \frac{\binom{m}{j}\binom{n-m}{k-j}}{\binom{n}{k}}.$$

This probability distribution is called hyper-geometric.

Usually the total m of flawed items in unknown. A quality control procedure might be as follows. Some k items are sampled from a party of size n. If the quantity X of flawed items is below a limit c, $X \leqslant c$, the party is accepted. Parameters n and c should be chosen to minimize probabilities of wrong decisions.

Example 17: Testing a hypothesis on a probability of an event. Assume someone wants to verify if the probability of an event A equals p. Is it true that zero pocket in roulette appears with probability $1/37$? Is it true that a birth of a boy is as probable as birth of a girl, i.e. $p_{\text{boy}} = p_{\text{girl}} = 0.5$? After conducting n observation in identical conditions (what makes the outcomes of observations independent) he get a random number X of occurrences of that event. From Example 5 we know that X has the binomial probability distribution with parameters n, p. Its mathematical expectation equals $\mathbf{M}X = np$. If the hypothesis is true, X should be close to np and by a similar reason the count $(n - X)$ of 'failures' should be close to $n(1-p) = nq$ with $q = 1-p$. Consider the χ^2-statistic (read: *chi-square*):

$$\chi^2 = \frac{(X - np)^2}{np} + \frac{(n - X - nq)^2}{nq}.$$

If the hypothesis on the probability of event A is true, the value of χ^2 is likely to be small (because the numerators on the rations are expected small). A little transformation

$$\begin{aligned}\frac{(X - np)^2}{np} + \frac{(n - X - nq)^2}{nq} &= \frac{(X - np)^2}{np} + \frac{(n - X - n(1-p))^2}{nq}\\ &= \frac{(X - np)^2}{n}\left(\frac{1}{p} + \frac{1}{q}\right)\\ &= \frac{(X - np)^2}{npq},\end{aligned}$$

So, one has to study the law of probability distribution for the output random variable $\chi^2 = g(X)$ with $g(u) = (u - np)^2/npq$.

When input variables are discrete, computations are mainly based on summation formula (1.2) for favorable cases.

Theorem 7. *Let X, Y, and Z be discrete (generally, dependent) random variables. Then*

$$\mathbf{P}(X = u) = \sum_{v,w} \mathbf{P}(X = u, Y = v, W = w), \tag{3.1}$$

$$\mathbf{P}(X = u, Y = v) = \sum_{w} \mathbf{P}((X = u, Y = v, W = w). \tag{3.2}$$

The proof is a direct reference to (1.2). The discrete probability distributions on the left hand sides of (3.1) and (3.2) are called marginal distributions of X, and of X and Y, respectively (relative to the joint probability distribution of all three random variables).

Theorem 8. *Let X and Y be integer non-negative independent random variables. Then*

$$\mathbf{P}(X + Y = n) = \sum_{k=0}^{n} \mathbf{P}(X = k)\mathbf{P}(Y = n - k). \tag{3.3}$$

Proof. Since both X and Y are non-negative, event $X + Y = n$ has a finite number of favorable cases: $\{X = 0, Y = n\}$, $\{X = 1, Y = n - 1\}$, …, $\{X = n, Y = 0\}$. Then,

$$\begin{aligned}
\mathbf{P}(X + Y = n) &= \sum_{k=0}^{n} \mathbf{P}(X = k, Y = n - k) && \text{(by (1.2))} \\
&= \sum_{k=0}^{n} \mathbf{P}(X = k)\mathbf{P}(Y = n - k) && \text{(by independence)}
\end{aligned}$$

The formula is proven. □

Example 18: Poisson process. Assume customers arrive randomly at an on-line store of an world-wide electronic commerce company. Assume further the postulates: 1) the probability of an arrival between times t and $t + h$ equals $\lambda h + o(h)$ where $o(h)$ is a term negligibly small ($o(h)/h \to 0$ as $h \to 0$), no arrivals occur with probability $1 - \lambda h + o(h)$; 2) the probability of two or more arrivals between times t and $t + h$ is $o(t)$; 3) arrivals between times t_1 and t_2, between t_2 and t_3, etc occur statistically independently.

Denote by $\mathcal{N}_t$ the arrivals count during the time-interval from 0 to t, $t > 0$, let $\mathcal{N}_0 = 0$. We have thus defined an infinite family of random variables. The random variables $\{\mathcal{N}_t; t \geqslant 0\}$ are called the Poisson process with intensity λ (this will be explained in a moment). Divide the time interval $(0, t)$ into n equal slots of length t/n. Then no arrivals during the interval are observed if and only if no arrivals are observed during each of the slots. Hence the probability of the event $\mathcal{N}_t = 0$ equals

$$(1 - \lambda(t/n) + o(t/n))^n$$

because of Postulates 1 and 3. As the number of slots n grows the probability tends to $e^{-\lambda t}$. Let k be a positive integer, and n be much greater than k. Consider an event $\mathcal{N}_t = k$. The probability that two or more events occur in the same time slot is $o(t/n)$ by Postulate 2. Then each time-slot may hold zero or one event independently of the other time-slots. We may use the formula for the binomial probability distribution for the probability of the event $\mathcal{N}_t = k$:

$$\binom{n}{k}(\lambda(t/n) + o(t/n))^k(1 - \lambda(t/n) + o(t/n))^{n-k}.$$

By virtue of the Poisson theorem, this probability tends to

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

as n tends to ∞, since the probability of a success $\lambda(t/n) + o(t/n)$ vanishes so that the product $n \times (\lambda(t/n) + o(t/n)) \to \lambda$. In effect, the counts $\mathcal{N}_t$, $t \geqslant 0$ have the Poisson probability distributions with parameter λt. By Postulate 3, the increments

$$\mathcal{N}_{t_1}, \quad \mathcal{N}_{t_2} - \mathcal{N}_{t_1}, \quad \ldots, \quad \mathcal{N}_{t_n} - \mathcal{N}_{t_{n-1}}$$

are independent random variables for all $n \geqslant 2$, $0 < t_1 < t_2 < \ldots < t_n$. A increment $\mathcal{N}_{t+h} - \mathcal{N}_t$ has the Poisson probability distribution with parameter λh and mathematical expectation λt. The intensity of arrivals is then

$$\lim_{h \to 0} \frac{\mathbf{M}(\mathcal{N}_{t+h} - \mathcal{N}_t)}{h} = \lambda.$$

Let $\{\mathcal{N}_t; t \geqslant 0\}$ be a Poisson process with intensity λ of arrivals from one country to the on-line store, let $\{\tilde{\mathcal{N}}_t; t \geqslant 0\}$ be a Poisson process with intensity μ of arrivals from another country. Let us find the probability distribution for the total number of arrivals during time t. Since $\mathcal{N}_t$ has the Poisson probability distribution with parameter λt and $\tilde{\mathcal{N}_t}$ has the Poisson probability distribution with parameter μt,

$$\begin{aligned}
\mathbf{P}(\mathcal{N}_t + \tilde{\mathcal{N}_t} = n) &= \sum_{k=0}^{n} \mathbf{P}(\mathcal{N}_t = k)\mathbf{P}(\tilde{\mathcal{N}_t} = n-k) && \text{by (3.3)} \\
&= \sum_{k=0}^{n} \frac{(\lambda t)^k}{k!} e^{-\lambda t} \frac{(\mu t)^{n-k}}{(n-k)!} e^{-\mu t} \\
&= \frac{e^{-(\lambda+\mu)t}}{n!} \sum_{k=0}^{n} \frac{n!}{k!(n-k)!} (\lambda t)^k (\mu t)^{n-k} \\
&= \frac{(\lambda+\mu)^n t^n}{n!} e^{-(\lambda+\mu)t} && \text{by Newton's binomial formula.}
\end{aligned}$$

The sum of two independent Poisson process is again a Poisson process with sum intensity. ■□■

Example 19: Connection between Poisson process and exponential probability distribution. Let $\{\mathcal{N}_t; t \geqslant 0\}$ be a Poisson process with intensity λ. We are going to study an interarrival interval. The aim of this example is to demonstrate many ideas in cooperation.

Let us calculate the probability that next arrival takes place between time t and $t+h$ given that an arrival occurred as instant s, $s < t$. In course of calculations we will see that the condition, which can be written as $\mathcal{N}_{s-\delta} \neq \mathcal{N}_s$ for sufficiently small δ, has probability zero. So, we have to define the conditional probability in question again. We will use the following definition of the conditional probability of an event A given an arrival at time s:

$$\mathbf{P}(A \mid \text{arrival at time } s) = \lim_{\substack{\delta \to 0, \\ \delta > 0}} \frac{\mathbf{P}(A \cap \{\mathcal{N}_{s-\delta} \neq \mathcal{N}_s\})}{\mathbf{P}(\mathcal{N}_{s-\delta} \neq \mathcal{N}_s)}.$$

We have

$$\begin{aligned}
\mathbf{P}(\mathcal{N}_{s-\delta} \neq \mathcal{N}_s) &= 1 - \mathbf{P}(\mathcal{N}_s - \mathcal{N}_{s-\delta} = 0) \\
&= 1 - e^{-\lambda\delta}.
\end{aligned}$$

The first arrival after time s occurs between times t and $t+h$ if and only if $\mathcal{N}_s = \mathcal{N}_t$, $\mathcal{N}_{t+h} > \mathcal{N}_t$. Hence, the probability in the numerator of the ratio for the conditional probability is

$$\mathbf{P}(\mathcal{N}_s - \mathcal{N}_{s-\delta} > 0, \mathcal{N}_t - \mathcal{N}_s = 0, \mathcal{N}_{t+h} - \mathcal{N}_t > 0) \\ = (1 - e^{-\lambda\delta}) \cdot e^{-\lambda(t-s)} \cdot (1 - e^{-\lambda h}).$$

Then the ratio is

$$\frac{(1 - e^{-\lambda\delta}) \cdot e^{-\lambda(t-s)} \cdot (1 - e^{-\lambda h})}{(1 - e^{-\lambda\delta})} = e^{-\lambda(t-s)} \cdot (1 - e^{-\lambda h})$$

Denote the interarrival time by T, then the probability

$$\mathbf{P}(u < T < u + h) = e^{-\lambda u} \cdot (1 - e^{-\lambda h}), \qquad u = t - s$$

and the probability density of T is the exponential density $\lambda e^{-\lambda u}$. ■□■

Let us now consider continuous variables. Let random variable X be given together with its probability density $p(u)$, and a function $g(u)$ such that $g'(u) \neq 0$ in an interval $[u_1, u_2)$ (hence either $g'(u) > 0$ or $g'(u) < 0$ everywhere in $[u_1, u_2)$). Set $v_1 = g(u_1)$, $v_2 = g(u_2)$, let $u = h(v)$ is the inverse function of $g(u)$. It is necessarily a one-to-one mapping of $[v_1, v_2)$ if $h'(v) > 0$ or $(v_2, v_1]$ if $h'(v) < 0$, onto $[u_1, u_2)$. Denote by $\tilde{p}(u)$ the probability density of $Y = g(X)$. Then, by the change of variable in integration,

$$\mathbf{P}(u_1 < X < u_2) = \int\limits_{u_1}^{u_2} p(u)\, du = \int\limits_{v_1}^{v_2} h'(v) p(h(v))\, dv. \tag{3.4}$$

If $h'(v) > 0$ then $v_1 < v_2$. The integral in the right-hand side of (3.4) equals the probability $\mathbf{P}(v_1 < Y < v_2)$. But, if $h'(v) < 0$, then $v_2 < v_1$, and

$$\int\limits_{v_1}^{v_2} h'(v) p(h(v))\, dv = -\int\limits_{v_2}^{v_1} h'(v) p(h(v))\, dv = \int\limits_{v_2}^{v_1} |h'(v)|\, p(h(v))\, dv.$$

The right-hand side of the formula above is $\mathbf{P}(v_2 < Y < v_1)$. Finally, the probability density of $Y = g(X)$ should be (see (2.5)).

$$\tilde{p}(v) = |h'(v)|\, p(h(v)). \tag{3.5}$$

Example 20: Linear function of a Gaussian random variable. Let X have a normal probability distribution (2.7), and $g(u) = \alpha u + \beta$.

Then $h(v) = (v+\beta)/\alpha$, $h'(v) = 1/\alpha$. Then, by virtue of (3.5) and (2.7),

$$\begin{aligned}\tilde{p}(v) =& \frac{1}{|\alpha|} p\Big(\frac{v+\beta}{\alpha}\Big)\\ =& \frac{1}{(|\alpha|\sigma)\sqrt{2\pi}}\, e^{-(u-a\alpha-\beta)^2/2(\alpha\sigma)^2},\end{aligned}$$

i.e. $\alpha X + \beta$ has the normal probability distribution with parameters $\alpha a + \beta$ and $(\alpha\sigma)^2$. ■□■

Theorem 9. *Let X, Y, and Z be (dependent) random variables with joint probability density $p(u, v, w)$. Then the marginal densities of X, and of X and Y are*

$$p_1(u) = \iint\limits_{\substack{-\infty<v<\infty\\ -\infty w<\infty}} p(u,v,w)\, dvdw, \tag{3.6}$$

$$p_{1,2}(u,v) = \int\limits_{-\infty}^{\infty} p(u,v,w)\, dw. \tag{3.7}$$

Proof. First,

$$\begin{aligned}&\mathbf{P}(u_1 < X < u_2, v_1 < Y < v_2)\\ &\qquad\qquad = \mathbf{P}(u_1 < X < u_2, v_1 < Y < v_2, -\infty < Z < \infty).\end{aligned}$$

Then,

$$\mathbf{P}(u_1 < X < u_2, v_1 < Y < v_2, -\infty < Z < \infty) = \iiint\limits_{\substack{u_1<u<u_2\\ v_1<v<v_2\\ -\infty<w<\infty}} p(u,v,w)\, dudvdw.$$

Turning to a repeated integration from the triple integral we get

$$\iiint\limits_{\substack{u_1<u<u_2\\ v_1<v<v_2\\ -\infty<w<\infty}} p(u,v,w)\, dudvdw = \iint\limits_{\substack{u_1<u<u_2\\ v_1<v<v_2}} \Big(\int\limits_{-\infty}^{\infty} p(u,v,w)\, dw\Big) dudv.$$

Comparison of the right-hand side of this equality to (2.14) proves (3.6). To prove (3.7), write

$$
\begin{aligned}
&\mathbf{P}(u_1 < X < u_2, v_1 < Y < v_2) \\
&\qquad = \mathbf{P}(u_1 < X < u_2, -\infty < Y < \infty, -\infty < Z < \infty) \\
&\qquad = \iiint\limits_{\substack{u_1<u<u_2\\ -\infty<v<\infty\\ -\infty<w<\infty}} p(u,v,w)\,dudvdw \\
&\qquad = \int\limits_{u_1}^{u_2} \Bigg(\iint\limits_{\substack{-\infty<v<\infty\\ -\infty<w<\infty}} p(u,v,w)\,dvdw \Bigg) du.
\end{aligned}
$$

□

Example 21: Marginals for planar Gaussian probability density. The marginal probability densities

$$\int\limits_{-\infty}^{\infty} p(u,v)\,du \quad \text{and} \quad \int\limits_{-\infty}^{\infty} p(u,v)\,du$$

from the planar normal probability density (2.17) are normal probability densities

$$\frac{1}{\sigma_1\sqrt{2\pi}}\,e^{-\frac{(u-a)^2}{2\sigma_1^2}} \quad \text{and} \quad \frac{1}{\sigma_2\sqrt{2\pi}}\,e^{-\frac{(v-b)^2}{2\sigma_2^2}}$$

correspondingly. ■□■

Theorem 10. *Let X and Y be continuous random variables with joint probability density $p(x,y)$, let functions $g_1(x,y)$ and $g_2(x,y)$ be continuous and one-to-one mapping. Assume also that continuous partial derivatives of g_1 and g_2 with respect to x and y exist. Functions g_1 and g_2 map a rectangle $u_1 < u < u_2$, $v_1 < v < v_2$ into a planar domain G. Let the functional determinant*

$$\begin{vmatrix} \dfrac{\partial g_1}{\partial x} & \dfrac{\partial g_1}{\partial y} \\ \dfrac{\partial g_2}{\partial x} & \dfrac{\partial g_2}{\partial y} \end{vmatrix} \neq 0$$

in the rectangle. Denote by $x = h_1(u_1, u_2)$ and $y = h_2(u_1, u_2)$ the inverse transform for $u = g_1(x, y)$ and $v = g_2(x, y)$,

$$J(x,y) = \begin{vmatrix} \dfrac{\partial h_1}{\partial u} & \dfrac{\partial h_1}{\partial v} \\ \dfrac{\partial h_2}{\partial u} & \dfrac{\partial h_2}{\partial v} \end{vmatrix}.$$

Then the joint probability density $\tilde{p}(u, v)$ of random variables $U = g_1(X, Y)$ and $V = g_2(X, Y)$ is

$$\tilde{p}(u, v) = p(h_1(u, v), h_2(u, v))|J(h_1(x, y), h_2(x, y))|. \tag{3.8}$$

Proof. The proof is based on a formula from multivariate calculus:

$$\mathbf{P}(x_1 < X < x_2, y_1 < Y < y_2) = \iint\limits_{\substack{x_1<x<x_2\\ y_1<y<y_2}} p(x,y)\,dxdy$$

$$= \iint\limits_{G} p(h_1(u,v), h_2(u,v))|J|\,dudv = \mathbf{P}((U,V) \in G).$$

Since the choice of x_1, x_2, y_1, and y_2 is arbitrary, we get (3.8). □

Theorems 9 and 10 let us obtain useful formulae for probability density for results of main arithmetic operations on continuous random variables.

Example 22: Probability density for a sum of two variables. Let X and Y have a joint probability density $p(x, y)$. Then the probability density of $X + Y$ is

$$p_{X+Y}(u) = \int\limits_{-\infty}^{\infty} p(x, u - x)\,dx \tag{3.9}$$

$$= \int\limits_{-\infty}^{\infty} p(u - y, y)\,dy. \tag{3.10}$$

Indeed, let $U = X$ and $V = X + Y$, then $X = U$ and $Y = V - U$. Here $h_1(u, v) = u$ and $h_2(u, v) = v - u$ and the functional determinant J is

$$\begin{vmatrix} 1 & 0 \\ -1 & 1 \end{vmatrix} = 1.$$

Hence the joint probability density of X and $X+Y$ is

$$p(u, v-u) \cdot 1 = p(u, v-u).$$

Now the desired probability density follows from (3.6). ■□■

Example 23: Purchase statistics. Let the purchase sum for one customer have the normal probability density (2.4). Assuming that purchase sums of n different customers are independent and identically distributed, what is the probability distribution for the total cost of the sold goods and what is the average purchase sum? Denote by X_1, X_2, …, X_n the purchase sums of the customers. Then the joint probability density of X_1 and X_2 is

$$p(u,v) = \frac{1}{\sigma^2 \cdot 2\pi} e^{-((u-a)^2+(v-a)^2)/2\sigma^2}.$$

Then the probability density of $X_1 + X_2$ is

$$p_2(v) = \int_{-\infty}^{\infty} p(x, v-x)\, dx$$

where

$$\begin{aligned} p(x, v-x) &= \frac{1}{\sigma^2 \cdot 2\pi} e^{-((x-a)^2+(v-x-a)^2)/2\sigma^2} \\ &= \frac{1}{\sigma^2 \cdot 2\pi} e^{-((x-a)^2+(x-a-(v-2a))^2)/2\sigma^2} \\ &= \frac{1}{\sigma^2 \cdot 2\pi} e^{-(2(x-a)^2+2(x-a)(v-2a)+(v-2a)^2)/2\sigma^2} \\ &= \frac{1}{\sigma^2 \cdot 2\pi} e^{-(2(x-a+(v-2a)/2)^2+(v-2a)^2/2)/2\sigma^2} \\ &= \frac{\sqrt{2}}{\sigma\sqrt{2\pi}} e^{-2(x-a+(v-2a)/2)^2/2\sigma^2} \frac{1}{\sigma\sqrt{2}\sqrt{2\pi}} e^{-(v-2a)^2/4\sigma^2} \end{aligned}$$

So,

$$\int_{-\infty}^{\infty} p(x, v-x)\, dx = \frac{1}{\sigma\sqrt{2}\sqrt{2\pi}} e^{-(v-2a)^2/4\sigma^2}$$

We conclude that $X_1 + X_2$ has a normal probability distribution with parameters $2a$ and $2\sigma^2$. Now, $X_1 + X_2$ and X_3 are independent and we can repeat computations to find out that $X_1 + X_2 + X_3$ has the normal probability density with parameters $3a$ and $3\sigma^3$. By induction, the total cost of sold goods $X_1 + X_2 + \ldots + X_n$ has the normal probability density

$$p_n(v) = \frac{1}{\sigma\sqrt{n}\sqrt{2\pi}} e^{-(v-na)^2/2n\sigma^2}.$$

Then the average purchase is

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

We may apply result from Example 20 with $\alpha = 1/n$ and $\beta = 0$ to get the probability density

$$\frac{\sqrt{n}}{\sigma\sqrt{2\pi}} e^{-n(v-a)^2/2\sigma^2}.$$

with the mathematical expectation a and variance σ^2/n. ■□■

Example 24: Chi-square distribution (χ^2-distribution). Let X_1, X_2, ..., X_n be independent with standard normal distribution ($a = 0$, $\sigma = 1$). Let $Y_n = X_1^2 + X_2^2 + \ldots + X_n^2$. First we'll find the density of probability distribution of X_1^2. Assume $t > 0$. Then

$$\begin{aligned}
\mathbf{P}(t - \delta < X_1^2 < t + \eta) &= \mathbf{P}(-\sqrt{t+\eta} < X_1 < -\sqrt{t-\delta}) \\
&\qquad + \mathbf{P}(\sqrt{t-\delta} < X_1 < \sqrt{t+\eta}) \\
&= \int\limits_{-\sqrt{t+\eta}}^{-\sqrt{t-\delta}} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}\, du + \int\limits_{\sqrt{t-\delta}}^{\sqrt{t+\eta}} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}\, du \\
&= 2 \int\limits_{\sqrt{t-\delta}}^{\sqrt{t+\eta}} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}\, du, \\
p_{X_1^2}(t) &= \frac{2}{\sqrt{2\pi}} e^{-(\sqrt{t})^2/2} \cdot \frac{1}{2\sqrt{t}} = \frac{1}{\sqrt{2t\pi}} e^{-t/2}
\end{aligned}$$

Using the Euler's Gamma function

$$\Gamma(a) = \int\limits_0^\infty x^{a-1} e^{-x}\, dx, \quad \Gamma(a+1) = a\Gamma(a), \quad \Gamma(1/2) = \sqrt{\pi},$$

one can write

$$p_{X_1^2}(t) = \frac{t^{-1/2}e^{-\frac{t}{2}}}{2^{1/2}\Gamma(1/2)}\,.$$

By induction one can prove that the probability density of Y_n is

$$p_{Y_n}(x) = \frac{x^{\frac{n}{2}-1}e^{-x/2}}{2^{n/2}\Gamma(n/2)}.$$

Assuming the formula for n, we can prove it for $n+1$:

$$\begin{aligned}
p_{Y_{n+1}}(x) &= \int\limits_0^x \frac{u^{\frac{n}{2}-1}e^{-u/2}}{2^{n/2}\Gamma(n/2)}\frac{(x-u)^{-\frac{1}{2}}e^{-(x-u)/2}}{2^{1/2}\Gamma(1/2)}\,du \\
&= \frac{1}{2^{(n+1)/2}\Gamma(n/2)\Gamma(1/2)}e^{-x/2}\int\limits_0^x u^{\frac{n}{2}-1}(x-u)^{-\frac{1}{2}}\,du \\
&= \frac{x^{\frac{n+1}{2}-1}}{2^{(n+1)/2}\Gamma(n/2)\Gamma(1/2)}e^{-x/2}B\Big(\frac{n}{2},\frac{1}{2}\Big) \\
&= \frac{x^{\frac{n+1}{2}-1}}{2^{(n+1)/2}\Gamma((n+1)/2)}e^{-x/2}.
\end{aligned}$$

The parameter n is called *degrees of freedom.* ■□■

Example 25: Probability density for a product of two variables. Let X and Y have a joint probability density $p(x,y)$. Then the probability density of $X \cdot Y$ is

$$\begin{aligned}
p_{X\cdot Y}(u) &= \int\limits_{-\infty}^{\infty} \frac{p(x,u/x)}{|u|}\,dx \\
&= \int\limits_0^{\infty} \frac{p(x,u/x)}{u}\,dx - \int\limits_{-\infty}^{0} \frac{p(x,u/x)}{u}\,dx. \qquad (3.11)
\end{aligned}$$

Really, let $U = X$ and $V = XY$, then $X = U$ and $Y = V/U$. Here $h_1(u,v) = u$ and $h_2(u,v) = v/u$ and the functional determinant J is

$$\begin{vmatrix} 1 & 0 \\ -v/u^2 & 1/u \end{vmatrix} = 1/u.$$

Hence the joint probability density of X and $X \cdot Y$ is

$$p(u, v/u)\frac{1}{|u|}.$$

Now the desired probability density follows from (3.6). ■□■

Example 26: Probability density for a ratio of two variables. Let X and Y have a joint probability density $p(x, y)$. Then the probability density of X/Y is

$$\begin{aligned} \tilde{p}_{X/Y}(u) &= \int\limits_{-\infty}^{\infty} |x|\, p(xu, x)\, dx \\ &= \int\limits_{0}^{\infty} xp(xu, x)\, dx - \int\limits_{-\infty}^{0} xp(xu, x)\, dx. \end{aligned}$$

Really, let $U = Y$ and $V = X/Y$, then $Y = U$ and $X = UV$. Here $h_1(u, v) = v$ and $h_2(u, v) = uv$ and the functional determinant J is

$$\begin{vmatrix} 0 & 1 \\ v & u \end{vmatrix} = -v.$$

Hence the joint probability density of X and $X \cdot Y$ is

$$|v| p(uv, v).$$

Now the desired probability density follows from (3.6). ■□■

Example 27: Student's t and Fisher's F. Next two continuous probability distribution play important role in statistics. If X is a normal random variable with $a = 0$ and $\sigma = 1$, $Y \geqslant 0$ is a continuous random variable independent of X and nY^2 has the χ^2-distribution with n degrees of freedom, then after certain efforts one obtains by (3.11) that $t = X/Y$ has the probability density

$$p(u) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi n}\Gamma\left(\frac{n}{2}\right)} (1 + u^2/n)^{-(n+1)/2}\,.$$

It is called the t-distribution with n degree of freedom.

Let random variables X and Y have χ^2-distribution with m and n degrees of freedom. Then the ration

$$F = \frac{X/m}{Y/n} = \frac{nX}{mY}$$

has a probability density

$$p(u) = \left(\frac{m}{n}\right)^{m/2} \frac{\Gamma((m+n)/2)}{\Gamma(m/2)\Gamma(n/2)} \cdot \frac{u^{m/2-1}}{(1+mu/n)^{(m+n)/2}}, \qquad u > 0.$$

This is called the F-distribution with m and n degrees of freedom. ■□■

Of course, theorem 10 doesn't cover all transformations important for practice. Let X_1, X_2, ..., X_n be independent identically distributed random variables. The *order statistics* are random variables $X_{1:n}$, $X_{2:n}$, ..., $X_{n:n}$ such that $X_{k:n}$ is the k-th in magnitude among X_1, X_2, ..., X_n. On other words, the order statistics are the variables X_1, X_2, ..., X_n sorted in ascending order. $X_{1:n}$ is the smallest, $X_{n:n}$ is the largest. The transformation carried out by sorting is not a one-to-one mapping, since $n!$ different permutations of same n values result in equal order statistics. So, Theorem 10 may not be applied to find the joint probability density for continuous random variables X_1, X_2, ..., X_n.

Example 28: Planning a market warehouse. The daily demand X at a city market is a random variable with known *cumulative distribution function* $F(u) = \mathbf{P}(X < u)$. The warehouse capacity should be greater than the largest demand during a year. By every next morning the warehouse is refilled with stocks to be ready for the new working day of the market. If the demand is greater than the warehouse capacity then sellers lose potential profits. Let X_1, X_2, ..., X_n be the demands for n successive working days ($n = 365$). They are assumed independent and identically distributed with X. The largest demand is then $Y = X_{n:n} = \max\{X_1, X_2, \ldots, X_n\}$. Knowing the cumulative distribution function $F_n(u) = \mathbf{P}(Y < u)$ is fruitful: choose the warehouse capacity C_α as the smallest value such that $F_n(C_\alpha) \geqslant 1 - \alpha$. Then the largest demand exceeds the warehouse capacity with probability

$$\mathbf{P}(Y \geqslant C_\alpha) = 1 - F_n(C_\alpha) \leqslant \alpha,$$

which means that the profit is lost not more often than in $\alpha \times 100$ % of years. One has:

$$F_n(u) = \mathbf{P}(X_{n:n} < u) = 1 - \mathbf{P}(X_1 < u, X_2 < u, \ldots, X_n < u) = 1 - (F(u))^n.$$

■□■

Lecture 4 Chebyshev's inequality and Laws of large numbers

Often the probability distribution of a random variable X is unknown. One may use statistical techniques to guess it, as it will be demonstrated in next lectures. Surprisingly, knowledge of just a few characteristics of X can be useful to give bounds on X or on certain probabilities of events generated by X. Moreover, such inexact bounds can be used to establish nontrivial facts about series of random variables.

The basic result here if the famous Chebyshev's inequality.

Theorem 11. *If X is a random variable with mathematical expectation $\mathbf{M}X = m$ and variance $\sigma^2 = \mathbf{var}\,X$. Then*

$$\mathbf{P}(|X - m| > t) \leqslant \frac{\sigma^2}{t^2}. \tag{4.1}$$

Proof. Assume additionally that X is continuous with probability density $p(u)$. This assumption simplifies the proof. The statement holds for arbitrary random variables with finite m and σ^2. We have

$$\begin{aligned}
\mathbf{var}\,X &= \int\limits_{-\infty}^{\infty} (u-m)^2\, p(u)\, du \\
&= \int\limits_{|u-m|>t} (u-m)^2\, p(u)\, du + \int\limits_{|u-m|\leqslant t} (u-m)^2\, p(u)\, du \\
&\geqslant \int\limits_{|u-m|>t} (u-m)^2\, p(u)\, du \\
&\geqslant \int\limits_{|u-m|>t} t^2\, p(u)\, du = t^2\, \mathbf{P}(|X-m| > t).
\end{aligned}$$

Dividing by t^2 we get the desired result. □

The meaning of the Chebyshev's inequality is in the ability to estimate the probability that the random variable X takes on a value from an interval

$(m-t, m+t)$. It is a symmetric interval about the mean value m of X. In general, an interval (u_1, u_2) is called a confidence interval for the random variable X with confident probability γ if $\mathbf{P}(u_1 < X < u_2) = \gamma$. Put $t = 3\sigma$ into (4.1). Then

$$\begin{aligned}\mathbf{P}(m-3\sigma \leqslant X \leqslant m+3\sigma) &= \mathbf{P}(|X-m| \leqslant 3\sigma) \\ &= 1 - \mathbf{P}(|X-m| > 3\sigma) \geqslant 1 - \frac{\sigma^2}{9\sigma^2} = 8/9,\end{aligned}$$

and $(m-3\sigma, m+3\sigma)$ is a confidence interval for X of significance level more than $8/9 = 0.88\ldots$, more that 88% observations of X belong to the interval. This is not a tight bound. If X has a uniform continuous probability distribution in an interval (a, b), then $m = (a+b)/2$ and $3\sigma = \sqrt{3}(b-a)/2 \approx 0.87(b-a)$, hence $m - 3\sigma < a$, $b < m + 3\sigma$ and the event $|X - m| < 3\sigma$ is the certain event. Its probability is 1. Now let X have a normal probability distribution with mathematical expectation a and variance σ^2, then

$$\begin{aligned}\mathbf{P}(|X-a| \leqslant 3\sigma) &= \int\limits_{-3\sigma}^{3\sigma} \frac{1}{\sigma\sqrt{2\pi}} e^{-(u-a)^2/2\sigma^2}\, du \\ &= \int\limits_{-3}^{3} \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\, dy \quad \text{by change of variable } y = (u-a)/\sigma \\ &= 0.99730\ldots \qquad \text{by numerical integration.}\end{aligned}$$

The Chebyshev's inequality is a convenient tool when exact computation of the probability is tedious.

Example 29: Estimating the probability of an event. How many trials are sufficient to estimate the probability p of an event A with confidence probability γ? The number X of occurrences of the event A in n independent trials has the binomial probability distribution with parameters n and p. We will estimate the unknown probability p be means of the relative frequency $\hat{p} = X/n$. Then the goal is to find a value t such that

$$\mathbf{P}(|\hat{p} - p| \leqslant t) \geqslant \gamma.$$

We have

$$\mathbf{M}\hat{p} = \frac{1}{n}\mathbf{M}X = \frac{np}{n} = p, \quad \mathbf{var}\,\hat{p} = \frac{1}{n^2}\mathbf{var}\,X = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n},$$

and the Chebyshev's inequality implies that

$$\mathbf{P}(|\hat{p} - p| \leqslant t) \geqslant 1 - \frac{p(1-p)}{t^2 n} \geqslant 1 - \frac{1}{4nt^2}\,. \tag{4.2}$$

So, n should be greater than $1/4t^2(1-\gamma)$. If, for instance, $\gamma = 0.95$ (95 %) and $t = 0.02$ then 12 500 trials are sufficient. ■□■

J. Bernoulli[1] in his posthumous book 'Ars conjectandi' (1713) formulated the following statement and gave the first proof to it.

Theorem 12 (Bernoulli's Law of large numbers)**.** *Let $\hat{p}_n$ be the relative frequency of an event A in n independent repeated trials, then:*

$$\lim_{n\to\infty} \mathbf{P}(|\hat{p}_n - p| \leqslant t) = 1$$

for all positive t.

Informally, any error in estimating the probability can be made as little probable as desired by choosing a sufficiently large number of trials. The proof is simple if the Chebyshev's inequality is at hand. Just send $n \to \infty$ in (4.2).

The first generalization of the Bernoulli's law of large numbers was done by Poisson.

Theorem 13. *Let the probability of the event A in the k-th independent trial equal p_k, $k = 1, 2, \ldots,$ and X_n if the number of occurrences of A in the first n trials. Then*

$$\lim_{n\to\infty} \mathbf{P}(|X/n - (p_1 + p_2 + \ldots + p_n)/n| \leqslant t) = 1$$

for all positive t.

Poisson's interpretation of his theorem concerned jury courts. Let the k-th member of the jury make right decision with probability p_k, then the proportion of right decisions in a jury with n members should be close to the average probability $(p_1 + p_2 + \ldots + p_n)/n$.

Attentive reader should recall the phenomenon of statistical stability discussed in the first lecture. The theorems of J. Bernoulli and Poisson look like a mathematical proof of the existence of statistical stability. Such

[1] Jacob Bernoulli (1655–1705) was a Swiss mathematicians, one of the Bernoullis. Famous for his contribution to probability theory, calculus, differential equations, and geometry

conclusion is illusive. We agreed to apply probability theory to random experiments which behave in a certain way. Thus, the laws of large numbers should be present in probability theory if the theory has anything to do with reality, and they can't make unstable experiments behave differently just because of a mathematical theory.

The name of the law reminds that the regularity in randomness becomes visible when the number of trials is large. This being said, the trials don't necessity need to be independent. Here we will give one simple but useful example which is rarely met in probability and statistics coursebooks.

Example 30: Sampling without replacement is representative. In the quality control example the items were picked for tests without replacement. The initial proportion of flawed items being $p = m/n$, what is the proportion $\hat{p} = X/k$ of flawed items in the sample? Here X has a hypergeometric probability with mathematical expectation $\mathbf{M}X = m/n = p$ and variance $\mathbf{var}\, X = km(n-m)(n-k)/n^2(n-1)$. By the Chebyshev's inequality,

$$\mathbf{P}(|\hat{p} - p| > t) \leqslant \frac{\mathbf{var}\,(X/k)}{t^2} = \frac{1}{k^2} \cdot \frac{km(n-m)(n-k)}{n^2(n-1)} \cdot \frac{1}{t^2}\,.$$

If m, n grow such that $m/n \to p$ and k also grows to infinity then the right-hand side of the inequality tends to 0, so any significant deviation of the observed proportion $\hat{p} = X/k$ from the actual proportion p becomes very little likely in a sample of large size k. ■□■

Now let us recall Example 23 where we had found that the average

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

of independent identically distributed random variables X_1, X_2, ..., X_n with normal probability density has in turn the normal probability density with the mathematical expectation a and variance σ^2/n. The confidence probability for $\overline{X}$ and the confidence interval $(a-t, a+t)$ equals

$$\int\limits_{-(a-t)}^{a+t} \frac{\sqrt{n}}{\sigma\sqrt{2\pi}} e^{-n(u-a)^2/2\sigma^2}\, du = \int\limits_{-t\sqrt{n}/\sigma}^{t\sqrt{n}/\sigma} \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\, dy.$$

The integral in the right-hand side tends to 1 as $n \to \infty$. It means that the any given error in estimating of the mathematical expectation value

m by means of the average $\overline{X}$ of normal random variables can be made as little probable as desired by choosing a sufficiently large number n of observations (measurements). It was Chebyshev[2] who had transferred this result to a wide class of general random variables.

Definition 11. *The law of large numbers holds for random variables X_1, X_2, ..., X_n, ..., with finite mathematical expectations $m_i = \mathbf{M}X_i$, $i = 1$, 2, ... if and only if for all $t > 0$*

$$\lim_{n\to\infty} \mathbf{P}\Big(\Big|\frac{1}{n}(X_1 + X_2 + \ldots + X_n) - \frac{1}{n}(m_1 + m_2 + \ldots + m_n)\Big| \leqslant t\Big) = 1.$$

Theorem 14 (Chebyshev's law of large numbers)**.** *Let X_1, X_2, ..., X_n, ... be independent random variables with arbitrary probability distributions. Let the mathematical expectations $m_i = \mathbf{M}X_i$, $i = 1, 2, \ldots$ exist and let the variances $\mathbf{var}\, X_i = \sigma_i^2$ be bounded by a constant C, $\sigma_i^2 < C$, for all $i = 1, 2, \ldots$. Then the law of large numbers holds for the random variables X_1, X_2, ..., X_n, ...*

Proof. The key to the proof is again the Chebyshev's inequality. We have

$$\begin{aligned}
\mathbf{M}\Big(\frac{1}{n}(X_1 + X_2 + \ldots + X_n)\Big) &= \frac{1}{n}(m_1 + m_2 + \ldots + m_n),\\
\mathbf{var}\,\frac{1}{n}(X_1 + X_2 + \ldots + X_n) &= \frac{1}{n^2}(\mathbf{var}\, X_1 + \mathbf{var}\, X_2 + \ldots + \mathbf{var}\, X_n)\\
&= \frac{1}{n^2}(\sigma_1^2 + \sigma_2^2 + \ldots + \sigma_n^2)\\
&\leqslant \frac{nC}{n^2} = \frac{C}{n}\,.
\end{aligned}$$

By Chebyshev's inequality,

$$\mathbf{P}\Big(\Big|\frac{1}{n}(X_1 + X_2 + \ldots + X_n) - \frac{1}{n}(m_1 + m_2 + \ldots + m_n)\Big| \leqslant t\Big) \geqslant 1 - \frac{C}{nt^2} \to 1$$

as $n \to \infty$. □

Now we are able to extend Example 23 by less restrictive assumptions that the probability distribution of X is not normal but the variance σ^2 is finite. The Chebyshev's law of large numbers states that the average purchase still stabilizes around the mathematical expectation $\mathbf{M}X$.

[2] Pafnuty Lvovich Chebyshev (1821–1894) was a Russian mathematician. His fields of interests were probability theory, approximation theory, and mechanics.

The top generalization of the Chebyshev's theorem belongs to Khinchine[3]. His theorem releases the requirement on the existence of variances at a cost of identical distributions.

Theorem 15. *If X_1, X_2, ..., X_n, ... are independent identically distributed random variables with finite mathematical expectation $m = \mathbf{M}X_1$ then*

$$\lim_{n\to\infty} \mathbf{P}(|(X_1 + X_2 + \ldots + X_n)/n - m| \leqslant t) = 1 \qquad \textit{for all } t > 0.$$

Example 31: Estimation of moments of a random variable. Let X be a random variable. Its moment $m^{(k)}$ of order k is defined as the mathematical expectation $\mathbf{M}(X^k)$. If $\mathbf{M}X^k$ exists then the sequence of averages

$$\overline{X^k} = \frac{1}{n}(X_1^2 + X_2^k + \ldots + X_n^2)$$

follows the law of large numbers, so for practical purposes

$$\frac{1}{n}(X_1^2 + X_2^k + \ldots + X_n^2) \approx m^{(k)}.$$

For example, since $\mathbf{var}\, X = m^{(2)} - (m^{(1)})^2$, one could compute $\overline{X}$ and $\overline{X^2}$ and use them to get an approximate value of the unknown $\sigma^2 = \mathbf{var}\, X$.

Recall the groom's ages example from the first lecture. We don't know the values taken on by the random values because the observed values were grouped into classes. Having intention to estimate the unknown mathematical expectation and variance we may substitute the middle points of the classes as 'typical values'; we also have to omit the cases when the age is unknown. Then we have $23\,919 - 698 = 23\,221$ observations in 2009, 655 observation gave approximately age 17, 4 786 observations age 22, etc. The observations in the class "> 60" will be replaced by age 65. The average age is, then,

$$\begin{aligned}(17 \cdot 655 + 22 \cdot 4686 + 27 \cdot 6587 + 32 \cdot 4330 + 37 \cdot 2374 + 42 \cdot 1507 \\ + 47 \cdot 1086 + 52 \cdot 693 + 57 \cdot 470 + 65 \cdot 733)/23221 = 32.009 \approx m^{(1)}.\end{aligned}$$

[3] Aleksandr Yakovlevich Khinchine (1894–1959) was a Soviet mathematician and a Correspondent Member of the Academy of Sciences of the USSR. His areas of research were probability theory, stochastic processes, real analysis, metric theory of functions, and number theory.

The average squared age is

$$(17^2 \cdot 655 + 22^2 \cdot 4686 + 27^2 \cdot 6587 + 32^2 \cdot 4330 + 37^2 \cdot 2374 + 42^2 \cdot 1507$$
$$+ 47^2 \cdot 1086 + 52^2 \cdot 693 + 57^2 \cdot 470 + 65^2 \cdot 733)/23221 = 1141.135 \approx m^{(2)}.$$

So, the unknown variance is

$$\sigma^2 \approx 141.135 - (32.009)^2 = 116.559.$$

A remark must be made here: the law of large numbers means that much more possible values of the average of the random variables lie close to the constant mean value than far from it, and we used experimental data to get just one of such typical values. But we can't guarantee that it *is* one of such values. A mistake is still possible. ■□■

If the mathematical expectations don't exist then the law of large numbers becomes inapplicable. Today classical laws of large numbers can be replaced by more general theorems of the same manner. Although random variables without mathematical expectation play important role in modern social sciences they are not covered in this concise course.

Unfortunately, the approach to the estimation of an unknown variance in the previous Example introduces a systematic error (vanishing with growth of the number of observations n), because

$$\mathbf{M}(\overline{X})^2 \neq (\mathbf{M}X)^2.$$

It is confirmed by the direct calculation:

$$\begin{aligned}\mathbf{M}(\overline{X})^2 &= \frac{1}{n^2}\mathbf{M}(X_1 + X_2 + \ldots + X_n)^2 \\ &= \frac{1}{n^2}\mathbf{M}(X_1^2 + X_2^2 + \ldots + X_n^2 + 2X_1X_2 + 2X_1X_3 + \ldots + 2X_{n-1}X_n) \\ &= \frac{1}{n^2}(nm^{(2)} + n(n-1)(m^{(1)})^2) = \frac{n-1}{n}(m^{(1)})^2 + \frac{m^{(2)}}{n}.\end{aligned}$$

In statistical parlance, $\overline{X^2} - (\overline{X})^2$ is called a *biased* estimator for the variance. Its bias is

$$\begin{aligned}\mathbf{M}(\overline{X^2} - (\overline{X})^2) - \sigma^2 &= \Big(m^{(2)} - \frac{n-1}{n}m^{(1)} - \frac{1}{n}m^{(2)}\Big) - \Big(m^{(2)} - (m^{(1)})^2\Big) \\ &= \frac{(m^{(1)})^2 - m^{(2)}}{n} = -\frac{\sigma^2}{n}\,.\end{aligned}$$

It's getting smaller as n grow, so the estimator is *asymptotically unbiased.*

Laws of large numbers describe in some sense the limiting behavior of sequences of averages. Inspired by the laws of large numbers is a notion of a *limit in probability.*

Definition 12. *A sequence of random variables X_1, X_2, ..., X_n, ... is said to* converge in probability *to a random variable X_0 if*

$$\lim_{n\to\infty} \mathbf{P}(|X_n - X_0| \leqslant t) = 1$$

for all $t > 0$.

Although the law of large numbers assumes a whole sequence X_1, X_2, ... is given, its application in data analysis deals with a finite number n of observed variables X_1, X_2, ..., X_n. Here the number n is selected such that the probability $\mathbf{P}(|X_n - X_0| \leqslant t) \geqslant \gamma$ for fixed $\gamma \approx 1$. The choice of n doesn't forbid realization of the opposite inequality $|X_m - X_0| > t$ for some larger values $m > n$ of the index.

A law of large numbers is a particular case of convergence in probability when X_n is the average of a sequence of variables and X_0 is a constant. Convergence in probability of random variables has many important properties of the ordinary convergence in calculus. We will need further next statement: if $\{X_n; n = 1, 2, \ldots\}$ converges in probability to a constant a and $\{Y_n; n = 1, 2, \ldots\}$ converges in probability to a constant b, and $g(u, v)$ is a continuous function at $u = a$ and $v = b$ then $\{g(X_n, Y_n); n = 1, 2, \ldots\}$ converges in probability to $g(a, b)$.

Example 32: Estimator $\overline{X^2} - (\overline{X})^2$ rehabilitated. Despite the estimator $\overline{X^2} - (\overline{X})^2$ of an unknown variance of X is biased, we will prove it converges in probability to the variance, so that the error in estimation vanishes (strictly speaking, the error converges to zero in probability). Let $g(u, v) = u - v^2$, then $g(u, v)$ is continuous and $g(m^{(2)}, m^{(1)}) = \sigma^2$. $\overline{X^2}$ converges in probability to $m^{(2)}$ by the law of large numbers, $\overline{X}$ converges in probability to $m^{(1)}$ again by the law of large numbers. So, $\overline{X^2} - (\overline{X})^2$ does converge in probability to σ^2. ■□■

Example 33: Estimator of warehouse capacity. In Example 28 we assumed that the cumulative distribution function of daily demand is known. Often this is not the case but one has collected observations for several years for the largest demand per year. Let Y_1, Y_2, ..., Y_r be the largest daily demands over r years. They are random variables with the same cumulative distribution function $F_n(u)$ (function $F_n(u)$ was introduced in Example 28). Let $Y_{1:r}$, $Y_{2:r}$, $\dots$, $Y_{r:r}$ be the order statistics from the variables Y_1, Y_2, $\dots$, Y_r (see page 51). For $0 < \gamma < 1$ let $[\gamma r]$ denote the largest integer less or equal to γr. The order statistic $Y_{[\gamma r]:r}$ is called a *sample quantile of level* γ. Put $\alpha = 1 - \gamma$. It turns out, $Y_{[\gamma n]:n}$ converges in probability to C_α as n grows if $F_n(u)$ is continuous at $u = C_\alpha$ and if an equation $F_n(C_\alpha) = 1 - \alpha$ has a unique solution. Let us prove this statement.

Let us assume first that Y_1, Y_2, $\dots$, Y_r have the uniform probability distribution in the interval $(0, 1)$. Pick numbers $0 < u_1 < u_2, \dots < u_r < 1$ and small strictly positive numbers δ_1, δ_2, $\dots$, δ_r. Inequalities

$$u_1 < Y_{1:r} < u_1 + \delta_1, \quad u_2 < Y_{2:r} < u_2 + \delta_2, \quad \dots, \quad u_r < Y_{r:r} < u_r + \delta_r$$

take place in $r!$ mutually exclusive cases obtainable by permutation of indices: either

$$u_1 < Y_1 < u_1 + \delta_1, \quad u_2 < Y_2 < u_2 + \delta_2, \quad \dots, \quad u_r < Y_r < u_r + \delta_r;$$

or

$$u_1 < Y_2 < u_1 + \delta_1, \quad u_2 < Y_1 < u_2 + \delta_2, \quad \dots, \quad u_r < Y_r < u_r + \delta_r;$$

etc. Each favorable case has probability $\delta_1 \delta_2 \dots \delta_r$ due to independence and formula (2.6) in Example 8. Hence the probability

$$\mathbf{P}(u_1 < Y_{1:r} < u_1 + \delta_1, u_2 < Y_{2:r} < u_2 + \delta_2, \dots, u_r < Y_{r:r} < u_r + \delta_r) \\ = r! \delta_1 \delta_2 \dots \delta_r,$$

the joint probability density of the order statistics $Y_{1:r}$, $Y_{2:r}$, $\dots$, $Y_{r:r}$ is

$$p(u_1, u_2, \dots, u_r) = r! \quad \text{for } 0 < u_1 < u_2 < \dots < u_r < 1.$$

The probability density equals zero when its arguments are not in ascending order. This probability density coincides with the joint probability density of ratios S_1/S_{r+1}, S_2/S_{r+1}, $\dots$, S_r/S_{r+1} of sums $S_1 = Z_1$, $S_2 = Z_1 + Z_2$, $S_3 = Z_1 + Z_2 + Z_3$, $\dots$, $S_{r+1} = Z_1 + Z_2 + \dots + Z_{r+1}$ of independent exponential random variables with parameter λ (verify this by generalizing Theorem 10) and

$$\mathbf{P}(|Y_{[\gamma r]:r} - C_\alpha| \leqslant t) = \mathbf{P}(|S_{[\gamma r]}/S_{r+1} - C_\alpha| \leqslant t).$$

Now, by the law of large numbers, $S_{r+1}/(r+1)$ and $S_{[\gamma r]}/[\gamma r]$ converge in probability to $\mathbf{M}Z_1 = 1/\lambda$. Hence,

$$\frac{S_{[\gamma r]}}{S_{r+1}} = \frac{S_{[\gamma r]}}{[\gamma r]} \cdot \frac{r+1}{S_{r+1}} \cdot \frac{[\gamma r]}{r+1}$$

converges in probability to γ as $r \to \infty$. The cumulative distribution function of the uniform distribution $(0,1)$ if $F_n(u) = u$ for $0 < u < 1$. In result, $Y_{[\gamma r]:r}$ converges in probability to C_α as $r \to \infty$.

Now let $F_n(u)$ be arbitrary continuous cumulative distribution function and U_1, U_2, ..., U_r be independent random variables with uniform distribution in $(0,1)$, $U_{1:r}$, $U_{2:r}$, ..., $U_{r:r}$ the corresponding order statistics. Let $F_n^{-1}(u)$ be the inverse function for $F_n(u)$ and by a direct computation $F_n^{-1}(U_1)$, $F_n^{-1}(U_2)$, ..., $F_n^{-1}(U_n)$ have the same probability distribution as Y_1, Y_2, ..., Y_r. Then

$$\mathbf{P}(|Y_{[\gamma r]:r} - C_\alpha| \leqslant t) = \mathbf{P}(|F_n^{-1}(U_{[\gamma r]:r}) - F_n^{-1}(\gamma)| \leqslant t) \to 1 \quad \text{as } r \to \infty$$

because $U_{[\gamma r]:r}$ converges in probability to γ. ■□■

Lecture 5 Gaussian law as an approximation. The central limit theorem

Let study more in-depth the Gaussian[1] law of probability distribution with a probability density

$$p(u) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(u-a)^2/2\sigma^2} . \tag{5.1}$$

The probability to fall into an interval (u_1, u_2) is given by the integral

$$\int\limits_{u_1}^{u_2} \frac{1}{\sigma\sqrt{2\pi}} e^{-(u-a)^2/2\sigma^2}\, du$$

which can't be expressed in terms of elementary functions in finite way (although *infinite* series expansions are used to evaluate it numerically). So the very first fact to learn about the *family* of Gaussian densities is that by the change of a variable $v = (u - a)/\sigma$ computations are reduced to the *standard normal density* with the parameters $a = 0$ and $\sigma = 1$. Precisely, for any u_1 and $u_2 > u_1$ we have

$$\int\limits_{u_1}^{u_2} \frac{1}{\sigma\sqrt{2\pi}} e^{-(u-a)^2/2\sigma^2}\, du = \int\limits_{(u_1-a)/\sigma}^{(u_2-a)/\sigma} \frac{1}{\sqrt{2\pi}} e^{-v^2/2}\, dv.$$

Recalling the Newton–Leibnitz formula of definite integration one may write

$$\int\limits_{(u_1-a)/\sigma}^{(u_2-a)/\sigma} \frac{1}{\sqrt{2\pi}} e^{-v^2/2}\, dv = F\Big(\frac{u_2 - a}{\sigma}\Big) - F\Big(\frac{u_1 - a}{\sigma}\Big)$$

[1] Karl Friedrich Gauß (1777–1855) was a German mathematician, whose impact on mathematics and whose contribution to number theory, algebra, analysis, statistics, differential geometry, celestial mechanics, geodesy, etc delivered him a title of *Princeps mathematicorum* (Latin: the Prince of mathematicians).

where $F(u)$ is any of the antiderivatives of the standard normal density $e^{-u^2/2}/\sqrt{2\pi}$. The common choices of the antiderivative are

$$F(u) = \Phi(u) = \int_{-\infty}^{u} \frac{1}{\sqrt{2\pi}} e^{-v^2/2}\, dv,$$

$$F(u) = \Phi_0(u) = \int_{0}^{u} \frac{1}{\sqrt{2\pi}} e^{-v^2/2}\, dv,$$

$$F(u) = \frac{1}{2} \operatorname{Erf}\left(\frac{u}{\sqrt{2}}\right)$$

where $\operatorname{Erf}(u)$ is the *error function*

$$\operatorname{Erf}(u) = \frac{2}{\sqrt{\pi}} \int_{0}^{u} e^{-v^2}\, dv.$$

The function $\Phi(u)$ is the cumulative distribution function for the standard normal density. It is has the following features:

$$\Phi(-\infty) = 0, \quad \Phi(\infty) = 1, \quad \Phi(0) = 0.5, \qquad \Phi(-u) = 1 - \Phi(u).$$

The function $\Phi_0(u)$ represents the probability assigned in accordance with the standard normal density to the interval $(0, u)$. Moreover,

$$\Phi(u) = 0.5 + \Phi(u), \quad \Phi_0(0) = 0, \quad \Phi_0(u) = 0.5, \quad \Phi_0(-u) = \Phi_0(u).$$

The functions $\Phi(u)$, $\Phi_0(u)$ were tabulated and the tables can be found in many manuals (see page 107). So, if X has a normal probability distribution with the density (5.1) then

$$\mathbf{P}(u < X < v) = \Phi\left(\frac{u-a}{\sigma}\right) - \Phi\left(\frac{u-a}{\sigma}\right) = \Phi_0\left(\frac{u-a}{\sigma}\right) - \Phi_0\left(\frac{u-a}{\sigma}\right).$$

Often we will be interested in confidence interval $(-z, z)$ for the standard normal random variable symmetric about zero. Let us demonstrate the procedure for the confidence probability of 0.9. We'd like to solve the equation

$$\int_{-z}^{z} \frac{1}{\sqrt{2\pi}} e^{-v^2/2}\, dv = 0.9$$

for positive z. The left-hand side is

$$\Phi_0(z) - \Phi_0(-z) = 2\Phi_0(z),$$

so we have $\Phi_0(z) = 0.45$. From Table 3, $z \approx 1.65$.

Figure 5.1 shows some frequently used intervals.

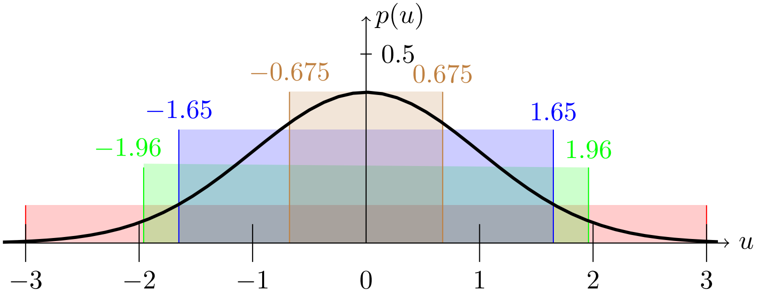

Figure 5.1. Symmetric confidence intervals for standard normal density: 50 % (brown), 90 % (light blue), 95 % (light green), 99.7 % (pink)

People know from practice that measurements usually are subject to errors. Mathematically, when someone is measuring a quantity m (e.g. a distance, an angle in geodesy, a weight, etc) he actually gets a number

$$Y = m + X,$$

where X is the error term. They observed that smaller errors occur more frequently than larger ones, that positive and negative errors of same magnitude are equally likely. So, X must have a symmetric probability distribution around 0 and its 'mean value' should be 0 as well. Different probability densities have been proposed for error terms (see Fig. 5). Gauss argued for his normal density because he had a kind of a proof for this formula from postulates he came up with. His main requirement was that the average of independent measurements Y_1, Y_2, ... should be 'the most probable' value for the unknown value of m.

Interestingly, the probability density named after him was known to mathematicians before Gauss. It first appeared as an approximation for the binomial probabilities in work of de Moivre[2].

[2] Abraham de Moivre (1667–1754) was a French mathematician. A Huguenot exiled to England, he was a friend of Isaac Newton, Edmond Halley, and James Stirling. Most famous for his work on probability theory and complex numbers.

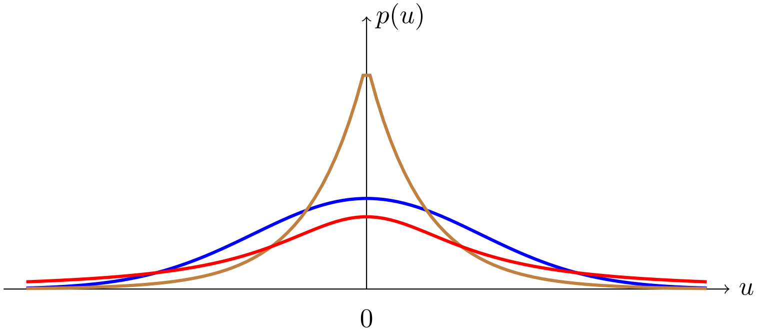

Figure 5.2. Different historical probability densities for errors: Gaussian density $e^{-u^2/2\sigma^2}/\sigma\sqrt{2\pi}$ (blue), two-sided Laplace's density $2e^{-\lambda\cdot|u|}/\lambda$ (brown), Cauchy's density $1/(\pi(1+u^2))$ (red)

Theorem 16 (The local theorem of de Moivre – Laplace)**.** *For a binomial probability distribution with parameters n and p and*

$$b(k;n,p) = \binom{n}{k} p^k q^{n-k}, \qquad q = 1-p,\ k = 0,1,\dots,n$$

we have

$$\frac{b(k;n,p)}{\dfrac{1}{\sqrt{npq}}\varphi(x_k)} \to 1 \qquad \text{as} \qquad k,n\to\infty$$

where

$$\varphi(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2} \qquad \text{and} \qquad x_k = \frac{k-np}{\sqrt{npq}}\ .$$

The proof to the theorem relies on the Stirling's approximate formula for the factorial function

$$n! \sim \sqrt{2\pi n}n^n e^n$$

and consists in direct computation of the limit.

Of course, the limit form of the theorem should be understood as follows: if n is large enough and p fixed, then the binomial probability $b(k;n,p)$ is approximately

$$\frac{1}{\sqrt{npq}}\varphi(x_k). \tag{5.2}$$

Table 5.1. Binomial probabilities approximated by local de Moivre – Laplace theorem for $p = 0.2$ and $n = 4$

k	x_k	$b(k;n,p)$	$\sqrt{np(1-p)}b(k;n,p)$	$\varphi(x_k)$
0	−1.0000	0.4096	0.3277	0.2420
1	0.2500	0.4096	0.3277	0.3867
2	1.5000	0.1536	0.1229	0.1295
3	2.7500	0.0256	0.0205	0.0091
4	4.0000	0.0016	0.0013	0.0001

The quality of approximation can be seen from Tables 5.1 and 5.2. ‘Large’ values of n are not really large.

Example 34: A telephone survey. In a telephone survey people chosen at random are called to be questioned. The probability a person refuses to answer to questions is 0.2. What is the probability 400 out of 500 agree to respond? Here $q = 0.2$, $p = 0.8$, $n = 500$, $np = 400$, $npq = 80$, $\sqrt{npq} = 4\sqrt{5} \approx 8.94$, $k = 400$, $x_{400} = (400 - 400)/4\sqrt{5} = 0$, and

$$\frac{1}{\sqrt{npq}}\varphi(x_{400}) = 0.044603\ldots.$$

The exact value of the probability of 400 successes is

$$\frac{500!}{400!100!}(0.8)^{400}(0.2)^{100} = 0.044564\ldots$$

The error is 0.000039 which is 0.088 % of the true probability. But the number of arithmetic operations needs to calculate the answer from (5.2) is much smaller! ■□■

From the local de Moivre – Laplace theorem we find that the probability of exactly $[np]$ successes approximately

$$\frac{1}{\sqrt{2\pi npq}}.$$

It vanishes as $n \to \infty$. It looks paradoxically that the probability of the mean value tends to zero, but it can easily resolved: the number of possible success counts becomes larger and larger, so each single probability necessarily diminishes.

Often the sum of binomial probabilities is needed. Let a, b be fixed numbers. Put $\Delta = (\sqrt{npq})^{-1}$. If we substitute $\Delta \cdot \varphi(x_k)$ into a sum

$$b(a;n,p) + b(a+1;n,p) + \ldots + b(b;n,p),$$

Table 5.2. Binomial probabilities approximated by the local de Moivre – Laplace theorem for $p = 0.2$ and $n = 4$

k	x_k	$b(k;n,p)$	$\sqrt{np(1-p)}b(k;n,p)$	$\varphi(x_k)$
0	−2.5000	0.0038	0.0076	0.0175
1	−2.0000	0.0236	0.0472	0.0540
2	−1.5000	0.0708	0.1417	0.1295
3	−1.0000	0.1358	0.2715	0.2420
4	−0.5000	0.1867	0.3734	0.3521
5	0.0000	0.1960	0.3920	0.3989
6	0.5000	0.1633	0.3267	0.3521
7	1.0000	0.1108	0.2217	0.2420
8	1.5000	0.0623	0.1247	0.1295
9	2.0000	0.0294	0.0589	0.0540
10	2.5000	0.0118	0.0236	0.0175
11	3.0000	0.0040	0.0080	0.0044
12	3.5000	0.0012	0.0023	0.0009
13	4.0000	0.0003	0.0006	0.0001
14	4.5000	0.0001	0.0001	0.00002
15	5.0000	0.00001	0.00002	0.00000

we obtain an integral sum

$$\sum_{k=a}^{b} \Delta \cdot \varphi(x_k)$$

for the definite integral

$$\int_{c_1}^{c_2} \varphi(u)\,du, \quad c_1 = \frac{a - np}{\sqrt{np(1-p)}}, \quad c_2 = \frac{b - np}{\sqrt{np(1-p)}}.$$

This leads to the famous de Moivre – Laplace result.

Theorem 17 (Integral theorem of de Moivre – Laplace)**.** *As* $n \to \infty$

$$\sum_{k=a}^{b} b(k;n,p) \to \int_{c_1}^{c_2} \varphi(u)\,du$$

where

$$c_1 = \frac{a - np}{\sqrt{np(1-p)}}, \qquad c_2 = \frac{b - np}{\sqrt{np(1-p)}}.$$

Example 35: Estimating the probability of an event. This example continues Example 29. Using the Integral theorem of de Moivre – Laplace, inequality (4.2) can be changed into the equality

$$\mathbf{P}(|\hat{p}-p| \leqslant t) = \mathbf{P}(np - tn < X < np + nt) \approx 2\Phi_0\Big(t\sqrt{\frac{n}{pq}}\Big) \geqslant 2\Phi_0(2t\sqrt{n}).$$

For $\gamma = 0.95$ and $t = 0.02$ we have

$$\Phi_0(0.04\sqrt{n}) \geqslant 0.475$$

and

$$0.04\sqrt{n} \geqslant 1.96, \qquad n = 2402.$$

We see that a more accurate estimation of the probability makes a sufficient number of experiments five times smaller. ■□■

Example 36: A cloakroom capacity problem. A movie theatre is planned to have two cloakrooms of equal capacity. Assuming there are 1000 seats in the theatre and every person chooses between the cloakrooms independently of the others and equiprobably, what capacity should it be so that one of them overflows in no more than 5 % of cases?

Let us denote by N the capacity of one cloakroom. Let X denote the number of people choosing the first cloakroom. Then there's no overflow when

$$X \leqslant N, \quad 1000 - X \leqslant N$$

whence

$$1000 - N \leqslant X \leqslant N.$$

We have Bernoulli trials with $n = 1000$ and $p = 1/2$. By virtue of the Integral De Moivre—Laplace theorem the probability of this event is

$$\begin{aligned}\Phi_0((N-np)/\sqrt{npq}) - \Phi_0((1000-N-np)/\sqrt{npq})& \\ = \Phi_0((N-500)/5\sqrt{10}) - \Phi_0((500-N)/5\sqrt{10}).&\end{aligned}$$

Obviously, $N \geqslant 500$, so $N - 500 \geqslant 0$ and $500 - N \leqslant 0$. Thus, the desired probability is

$$2\Phi_0((N-500)/5\sqrt{10}) \geqslant 0.95.$$

From the table,

$$\frac{N-500}{5\sqrt{10}} \geqslant 1.96$$

and $N > 530.99$, so $N = 531$. ■□■

Example 37: Testing a hypothesis on the probability of an event. Let us continue Example 17. Since X has a binomial probability distribution with parameters n and p, the χ^2-statistic has a probability distribution approximately as a squared standard normal variable, i.e. a χ^2-distribution with one degree of freedom. The Bernoulli's Law of large numbers tells us that typical values of χ^2 should be small. It means that large values of the χ^2-statistic give evidence against the hypothesized value p. For instance, 95 % observations of a random variable with χ^2-distribution with 1 degree of freedom lie in a interval between 0.0 and 3.841. ■□■

The Integral theorem of de Moivre and Laplace was the source of research efforts aimed to understand condition when the Gaussian probability distribution arise. Let us cite just a few later results in this direction.

Theorem 18 (Central limit theorem for independent identically distributed random variables)**.** *Let X_1, X_2, … be independent identically distributed random variables with finite expectation $m = \mathbf{M}X_1$ and variance $\sigma^2 = \mathbf{var}\,X_1$. Then for any $u_1 < u_2$*

$$\lim_{n\to\infty} \mathbf{P}\Big(u_1 < \frac{X_1 + X_2 + \ldots + X_n - nm}{\sigma\sqrt{n}} < u_2\Big) = \Phi(u_2) - \Phi(u_1).$$

Recall that $\Phi(x)$ is the cumulative distribution function of the standard normal distribution. Then Theorem 18 can be stated as follows: under the above-mentioned hypotheses the ratio

$$\frac{S_n - n\mathbf{M}X_1}{\sqrt{\mathbf{var}\,S_n}} \qquad \text{with } S_n = X_1 + X_2 + \ldots + X_n$$

has a probability distribution close to the standard normal distribution when n is large.

In Lecture 2 we saw that a random variable X_n with a binomial probability distribution came to out attention as the number of successes in n independent experiments. Denoting by Y_k the number of successes in k-th trial (0 or 1) we have $X_n = Y_1 + Y_2 + \ldots + Y_n$, $\mathbf{M}Y_k = p$, $\mathbf{var}\,Y_k = pq$ and variables Y_1, Y_2, …, Y_n are independent. Then the theorem of de Moivre and Laplace is a particular case of Theorem 18.

Theorem 18 plays important role in probability theory and mathematical statistics as well as in wide applications in technical and social sciences.

Example 38: Measurement errors. Consider a physical measurements with an error uniformly distributed in $(-1, 1)$ in some convenient units. If we have n measurements, what is the probability that the average value deviates from the true value by less than given limit $\delta > 0$?

Let

$$X_j = m + Y_j,$$

where X_j is the result of the j-th measurements, m the true value, Y_j the error, $-1 < Y_j < 1$. Then

$$\mathbf{M}Y_j = \int\limits_{-1}^{1} \frac{u}{2}\,du = 0, \quad \mathbf{var}\,Y_j = \mathbf{M}Y_j^2 = \int\limits_{-1}^{1} \frac{u^2}{2}\,du = \frac{1}{3},$$

the average value from n measurements is

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n} = \frac{mn + Y_1 + Y_2 + \ldots + Y_n}{n} = \\ = m + \frac{Y_1 + Y_2 + \ldots + Y_n}{n}.$$

By virtue of Theorem 18 we get

$$\begin{aligned} \mathbf{P}(|\overline{X} - m| < \delta) &= \mathbf{P}\Big(\Big|m + \frac{Y_1 + Y_2 + \ldots + Y_n}{n} - m\Big| < \delta\Big) \\ &= \mathbf{P}\Big(-\delta < \frac{Y_1 + Y_2 + \ldots + Y_n}{n} < \delta\Big) \\ &= \mathbf{P}\Big(-\delta\sqrt{3n} < \frac{Y_1 + Y_2 + \ldots + Y_n}{\sqrt{n/3}} < \delta\sqrt{3n}\Big) \\ &\approx \Phi_0(\delta\sqrt{3n}) - \Phi_0(-\delta\sqrt{3n}) = 2\Phi_0(\delta\sqrt{3n}). \end{aligned}$$

If, for instance, $n = 25$ and $\delta = 0.2$ then $2\Phi_0(\delta\sqrt{3n}) \approx 0.92$. ■□■

Example 39: Asymptotic normality of sample moments. Let X_1, $X_2, \ldots$ are independent identically distributed random variables. If $m^{(k)} = \mathbf{M}(X_1^k)$ and $m^{(2k)} = \mathbf{M}(X_1^{2k})$ exist then $\mathbf{var}\,(X_1^k) = \mathbf{M}(X^{2k}) - (\mathbf{M}X_1^k)^2$ and

$$\frac{\overline{X^k} - m^{(2k)}}{\sqrt{(m^{(2k)} - (m^{(k)})^2)n}}$$

is asymptotically normal. ■□■

Yet the power of asymptotic analysis of distributions of sums of a large number of random variables is collected in a theorem which doesn't assume identical distribution of terms. Such a theorem can be applied in many situations when a total of relatively small contributions of different nature make a total quantity with a normal probability distribution. In practice we will apply such a theorem when only mathematical expectations and variances of the small terms are known and not the complete probability distributions. We approximate the probability distribution for the sum by adjusting the mathematical distribution and variance for the sum only.

Theorem 19 (Lindeberg)**.** *Let* X_1*,* X_2*, …* *be independent random variables with mathematical expectations* $m_1 = \mathbf{M}X_1$*,* $m_2 = \mathbf{M}X_2$*, …, variances* $\sigma_1^2 = \mathbf{var}\, X_1$*,* $\sigma_2^2 = \mathbf{var}\, X_2$*, …* *and let* $B_n^2 = \sigma_1^2 + \sigma_2^2 + \ldots + \sigma^2 = \mathbf{var}\,(X_1 + X_2 + \ldots + X_n)$*. If, moreover, for all constant* $\tau > 0$ *we have*

$$\lim_{n\to\infty} \frac{1}{B_n^2} \sum_{k=1}^{n} \mathbf{M}((X_k - m_k)^2 \cdot Y_k(\tau)) = 0$$

where

$$Y_k(\tau) = \begin{cases} 1 & \textit{if } |X_k - m_k| > \tau B_k \\ 0 & \textit{otherwise} \end{cases}$$

then as $n \to \infty$ *uniformly in* u

$$\mathbf{P}\Big(\frac{1}{B_n} \sum_{k=1}^{n} (X_k - m_k) < u\Big) \to \int\limits_{-\infty}^{u} \frac{1}{\sqrt{2\pi}} e^{-v^2/2}\, dv.$$

The meaning of the Lindeberg's theorem is that

$$\frac{(X_1 - m_1) + (X_2 - m_2) + \ldots + (X_n - m_n)}{B_n}$$

has approximately the standard normal probability distribution.

Example 40: Stock pricing model. Let us look again at a financial market of Example 12. Assume that the returns R_1, R_2, …, R_t, … over several days are independent random variables. The price on day t is

$$S_t = S_0 \cdot R_1 \cdot S_2 \times \ldots \times S_t.$$

The logarithm of the price is the sum of logarithms of the returns:

$$\ln S_t = \ln S_0 + \ln R_1 + \ln S_2 + \ldots + \ln S_t.$$

If mathematical expectations $m_1 = \mathbf{M}\ln R_1$, $m_2\mathbf{M}\ln R_2$, $\ldots$ and variances $\sigma_1^2 = \textbf{var}\ \ln R_1$, $\sigma_2^2 = \textbf{var}\ \ln R_2$, $\ldots$ are known then the probability distribution of $\ln S_t$ should be approximately normal with mathematical expectation $m = m_1 + m_2 + \ldots + m_t$ and variance $\sigma^2 = \sigma_1^2 + \sigma_2^2 + \ldots + \sigma_t^2$. The probability distribution of the price S_t is called log-normal and has a probability density

$$p(u) = \frac{1}{u\sigma\sqrt{2\pi}} e^{-(\ln u - m)^2/2\sigma^2}, \qquad u > 0.$$

Example 41: "Revealing the mysteries of his technique". One shouldn't believe that the Gaussian probability distribution is a universal approximation tool, not even in case of summation. Here is an example proposed by S.D. Poisson (see the footnote on p. 23). Let X_1, X_2, $\ldots$, X_n be independent random variables with identical probability density of

$$p(u) = \frac{1}{\pi(1+u^2)}.$$

One is encouraged to verify by means of formula (3.10) that $X_1 + X_2$ has a probability density

$$p_2(u) = \frac{2}{\pi(4+u^2)},$$

$X_1 + X_2 + X_3$ has a probability density

$$p_3(u) = \frac{3}{\pi(9+u^2)},$$

and, generally, $X_1 + X_2 + \ldots + X_n$ has a probability density

$$p_n(u) = \frac{n}{\pi(n^2+u^2)}.$$

Then by virtue of formula (3.5), the average of n values

$$\frac{X_1 + X_2 + \ldots + X_n}{n}$$

has the same probability density $p(u)$ as every single term X_1, X_2, $\ldots$. Therefore, if the errors are distributed according to the density $p(u)$ then averaging doesn't reduce the total error.

There is a branch of studies in probability theory which investigates what limiting probability distortions may ever exist.

Lecture 6 Markov chains

A large number of working models in technical and social sciences are so-called Markov models. Imagine a process evolving in some state space in ways that its future after any time t depends (statistically) only on the process state at time t but not on the past before time t. In physics this property characterizes *dynamic systems* whose future is *determined* by the present but not by the past. Using a more high-end mathematics one can convert many processes into Markov ones by selection of an extended state space. Let us assume that the time is discrete, the discrete time variable is $n = 0, 1, \ldots,$ the state space is discrete (either finite or infinite denumerable). Denote by X_n the process state at time n.

Definition 13. *A sequence $X_0, X_1, \ldots, X_n, \ldots$ of discrete random variables is called a Markov chain*[1] *if for all values $a_0, a_1, \ldots, a_n$ such that*

$$\mathbf{P}(X_0 = a_0, X_1 = a_1, \ldots, X_{n-1} = a_{n-1}) > 0$$

one has

$$\mathbf{P}(X_n = a_n | X_0 = a_0, X_1 = a_1, \ldots, X_{n-1} = a_{n-1}) \\ = \mathbf{P}(X_n = a_n | X_{n-1} = a_{n-1}). \quad (6.1)$$

Equality (6.1) is called the *Markov property*. Denote

$$p_n(a, a') = \mathbf{P}(X_n = a' | X_{n-1} = a).$$

If functions $p_n(a, a') = p(a, a')$ are independent of n the Markov chain is called homogeneous (time-homogeneous). Then $p(a, a')$ is called the (one-step) transition probability of the Markov chain. The Markov property means exactly that the 'past' of the process up to time n affects its 'future' only through the 'present'. The probability distribution $p_0(a) = \mathbf{P}(X_0 = a)$ is called the *initial probability distribution* of the Markov chain. For the

[1] Andreĭ Andreevich Markov (1856–1922) was a Russian mathematician and a Member of Russian Imperial Academy of Sciences. He is famous for his works in probability theory, mathematical analysis, and number theory.

sake of simplicity we shall assume that the possible values of the random variables X_n are non-negative integers 0, 1, … for all $n = 0, 1, \dots$, possibly not greater than some integer constant M. If there's no such constant, put $M = \infty$. If $M < \infty$ the Markov chain is called finite, otherwise it is called denumerable (or countable). The Markov property (6.1) is the simplest type of statistical dependence next to pure independence: recalling the multiplication formula for probabilities (1.8) for several events we get

$$\begin{aligned}
&\mathbf{P}(X_0 = a_0, X_1 = a_1, \dots, X_n = a_n) \\
&= \mathbf{P}(X_0 = a_0) \times \mathbf{P}(X_1 = a_1 | X_0 = a_0) \times \mathbf{P}(X_2 = a_2 | X_0 = a_0, X_1 = a_1) \times \\
&\qquad \dots \times \mathbf{P}(X_n = a_n | X_0 = a_0, X_1 = a_1, \dots, X_{n-1} = a_{n-1}) \\
&\qquad\qquad = p_0(a_0) p_1(a_0, a_1) p_2(a_2, a_1) \cdots p_n(a_{n-1}, a_n).
\end{aligned}$$

This result also means that to define a time-homogeneous Markov chain with M states $\{u_1, u_2, \dots\}$ we need an initial probability distribution

$$p_0(u_k) \geqslant 0, \qquad 1 \leqslant k \leqslant M, \qquad \sum_{k=1}^{M} p_0(u_k) = 1$$

and a *matrix of transition probabilities*

$$P = \begin{pmatrix} p(u_1, u_1) & p(u_1, u_2) & \dots, & p(u_1, u_M) \\ p(u_2, u_1) & p(u_2, u_2) & \dots, & p(u_2, u_M) \\ \vdots & \vdots & \ddots & \vdots \\ p(u_M, u_1) & p(u_M, u_2) & \dots, & p(u_M, u_M) \end{pmatrix},$$

$$\sum_{k=1}^{M} p(u_j, u_k) = 1, \quad j = 1, 2, \dots, M.$$

The matrix is infinite if $M = \infty$ and there is no last column or last row in it. In this lecture we will deal only with time-homogeneous Markov chains.

Denote by

$$\begin{aligned}
p(n; a_1, a_2) &= \mathbf{P}(X_n = a_2 | X_0 = a_1) \\
&= \mathbf{P}(X_{n+m} = a_2 | X_m = a_1), \quad m = 1, 2, \dots
\end{aligned}$$

the probability of a transition from state a_1 to state a_2 in n steps, and

$$p(n; a) = \mathbf{P}(X_n = a)$$

for the probability if finding the process at state a at time n. Also, let us introduce a square matrix

$$P^{(n)} = (p(n; u_j, u_k) \colon 1 \leqslant j, k \leqslant M)$$

and a row-vector

$$\Pi^{(n)} = (p(n; u_j) \colon 1 \leqslant j \leqslant M).$$

As the following theorem demonstrates, these probabilities can be found fast by means of matrix algebra.

Theorem 20 (Kolmogorov–Chapman equations)**.** *The following equation holds:*

$$p(m + n; u_j, u_k) = \sum_{l=1}^{M} p(m; u_j, u_l) p(n; u_l, u_k), \tag{6.2}$$

or, in a matrix form,

$$P^{(m+n)} = P^{(m)} \cdot P^{(n)}. \tag{6.3}$$

Since $P^{(1)} = P$, we have $P^{(n)} = P^n$. Furthermore,

$$\Pi^{(m+n)} = \Pi^{(m)} \cdot P^{(n)}.$$

Proof. Simply, (6.2) follows from the law of total probability and the multiplication theorem:

$$\begin{aligned}
p(m + n; u_j, u_k) &= \mathbf{P}(X_{m+n} = u_k \mid X_0 = u_j) \\
&= \sum_{l=1}^{M} \mathbf{P}(X_{m+n} = u_j, X_m = u_l \mid X_0 = u_j) && \text{by (1.2)} \\
&= \sum_{l=1}^{M} \mathbf{P}(X_m = u_l \mid X_0 = u_j) \\
&\qquad \times \mathbf{P}(X_{m+n} = u_k \mid X_0 = u_j, X_m = u_l) && \text{by (1.7)} \\
&= \sum_{l=1}^{M} \mathbf{P}(X_m = u_l \mid X_0 = u_j) \\
&\qquad \times \mathbf{P}(X_{m+n} = u_k \mid X_m = u_l) && \text{by (6.1)} \\
&= \sum_{l=1}^{M} p(m; u_j, u_l) p(n; u_l, u_k).
\end{aligned}$$

The rest is about notation. Notice that (1.2) can be used when M is finite. Otherwise, more assumptions should be posed on probability $\mathbf{P}(\cdot)$, such as (1.2) holding true for a denumerable set of mutually exclusive events. □

In one important case the matrices P^n become little varying for sufficiently large n. Then the row vector Π^n of probabilities can be evaluated approximately much quicker. We need one more definition to formulate the corresponding result.

Definition 14. *Non-negative real numbers π_j, $k = 1, \ldots, M$ are called a* stationary probability distribution *for the Markov chain $\{X_n; n = 0, 1, \ldots\}$ when*

$$\pi_j = \sum_{k=1}^{M} \pi_k p(u_k, u_j), \quad j = 1, 2, \ldots, M; \qquad \sum_{j=1}^{M} \pi_j = 1. \tag{6.4}$$

If all $\pi_j > 0$, $j = 1, 2, \ldots, M$ then the stationary probability distribution is called ergodic.

The rationale for the name is as follows. Let us take the stationary probability distribution in place of the initial probability distribution for X_0, then

$$p(2; a) = \mathbf{P}(X_2 = a) = \sum_{k=1}^{M} p(1; u_k) p(u_k, a) = \pi_j,$$

and, in general, $p(n; u_j) = \pi_j$ and

$$\begin{aligned}\mathbf{P}(X_m = u_j, X_{m+1} = a_1, \ldots, X_{m+n} = a_n)& \\ = \pi_j p(u_j, a_1), p(a_2, a_2) \times \ldots \times p(a_{m+n-1}, a_{m+n})&.\end{aligned}$$

Theorem 21. *If $M < \infty$ and some power matrix P^n, $n \geqslant 1$ of the matrix P has all positive entries, then an unique ergodic distribution exists and*

$$\lim_{n \to \infty} p(n; u_j, u_k) = \pi_k, \qquad j = 1, 2, \ldots, M. \tag{6.5}$$

Even if the ergodic stationary probability distribution is not the initial probability distribution of the Markov chain, we can get an understanding of it from a frequentist's viewpoint. First, let us extend our toolset with conditional mathematical expectations. Let Y be a discrete random variables with possible values v_1, v_2, ... A conditional probability of an event A given $Y = v$ is, obviously,

$$\mathbf{P}(A \mid Y = v) = \frac{\mathbf{P}(A \cap \{Y = v\})}{\mathbf{P}(Y = v)}. \tag{6.6}$$

If X is another discrete random variable with possible values u_1, u_2, $\ldots$, then probabilities

$$\mathbf{P}(X = u \mid Y = v) = \frac{\mathbf{P}(X = u, Y = v)}{\mathbf{P}(Y = v)} \qquad \text{by (6.6)}$$

define the conditional probability distribution of X given $Y = v$. We can also define a suitable *conditional mathematical expectation*

$$\mathbf{M}(X \mid Y = v) = \sum_k u_k \mathbf{P}(X = u_k \mid Y = v).$$

Let A be a subset of state space of a Markov chain X_0, X_1, $\ldots$, X_n, $\ldots$. Let $I_n = 1$ if $X_n \in A$ and $I_n = 0$ otherwise. Here, finding the chain in a state from A is considered as a success and I_n represents the number of successes at time n. Consider

$$\nu_A(n) = \frac{I_1 + I_2 + \ldots + I_n}{n}$$

which is the fraction of the time spent by the physical system in the set A. It can also be considered as the relative frequency of visiting the set A. By the definition of the conditional mathematical expectation,

$$\mathbf{M}(I_t \mid X_0 = a) = \mathbf{P}(X_t \in A \mid X_1 = j) = \sum_{j=1}^{M} p(n; a, u_j).$$

Let us denote the right-hand side of the expression above by $p(n; a, A)$). Then we have

$$\mathbf{M}(\nu_A(n) \mid X_0 = a) = \frac{1}{n+1} \sum_{m=0}^{n} p(m; a, A).$$

It is known from calculus that if a sequence of real numbers $\{x_n; n = 0, 1, \ldots\}$ converges to a limit x, then

$$\lim_{n \to \infty} \frac{x_0 + x_1 + \ldots + x_n}{n+1} = x.$$

Hence if

$$\lim_{n \to \infty} p(n; a, u_j) = \pi_j,$$

then

$$\mathbf{M}(\nu_A(n)) = \sum_{j=1}^{M} \mathbf{M}(\nu_A(n) \mid X_0 = a)\mathbf{P}(X_0 = a) \to \pi_A, \quad \text{where } \pi_A = \sum_{u_j \in A} \pi_j.$$

A stronger statement is contained in the following theorem.

Theorem 22 (Law of large numbers for a finite Markov chain)**.** *If X_0, X_1, ..., X_n, ... if a finite a finite Markov chain with ergodic probability distribution π_j, $j = 1, 2, \ldots, M$, then*

$$\lim_{n\to\infty} \mathbf{P}(|\nu_A(n) - \pi_A| > \varepsilon) = 0$$

for every $\varepsilon > 0$ and every initial distribution.

Example 42: Occupational study [3]. Markov chains can be used to study intergenerational occupation mobility. Every working person belongs to one of a set of occupational classes (different researchers may choose different occupational classes in accordance with different socially relevant criteria). Then, the question is, how are occupational classes of fathers, grand-fathers, etc., and sons are related? Let the classes be 'non-manual labour', 'manual labour', and 'farming'. We have data from marriage-license applications for Marion County, Luisiana for periods of 1905–1912 and 1938–1941. The first sample contained 10 253 observations, and the second sample has 9 892 observations. Suppose we observe a random family: a grandfather, a father, a son, a grand-son, etc. Let $X_n = 1$ (non-manual), 2 (manual), 3 (farming) represent the occupational class in n-th generation, $n = 0, 1, \ldots$. We assume that the sequence makes a Markov chain. The transition probabilities can be estimated from these data by means of conditional frequencies. From the first set of data (1905–1912) we have

$$P = \begin{pmatrix} 0.594 & 0.396 & 0.009 \\ 0.211 & 0.782 & 0.007 \\ 0.252 & 0.641 & 0.108 \end{pmatrix}.$$

The matrix P has only strictly positive entries, so Theorem 21 can be applied here. The corresponding row-vector of stationary probabilities is found as the solution to system (6.4), which take the form

$$\pi_1 = 0.594\pi_1 + 0.211\pi_2 + 0.252\pi_3,$$

$$\begin{aligned}\pi_2 &= 0.396\pi_1 + 0.782\pi_2 + 0.641\pi_3,\\ \pi_3 &= 0.009\pi_1 + 0.007\pi_2 + 0.108\pi_3,\\ 1 &= \pi_1 + \pi_2 + \pi_3.\end{aligned}$$

The solution is

$$(0.343, 0.648, 0.09).$$

The actual fractions in each of the classes are

$$(0.310, 0.658, 0.034).$$

We observe that the actual number of farmers differs significantly from the predicted stationary one. This would suggest that over time the numbers of farmers should decrease.

In the second sample,

$$P = \begin{pmatrix} 0.622 & 0.357 & 0.003 \\ 0.274 & 0.721 & 0.005 \\ 0.265 & 0.694 & 0.042 \end{pmatrix},$$

the row-vector of stationary probabilities is

$$(0.420, 0.576, 0.04),$$

and the actual fractions are

$$(0.373, 0.616, 0.11).$$

As predicted, the fraction of farmers has significantly decreased. Also, one may notice that the transition matrices have close similar elements, so the experiment can be regarded as statistically stable. ■□■

Example 43: Moran inventory model. An inventory operates daily. In the morning of day n a supply of random size X_n arrives. The daily demand is constant and equals 1 if the morning stocks are present, and equals 0 if the stocks are absent. The inventory capacity is M. Let us assume that the variables X_1, X_2, ..., X_n, ... are independent identically distributed, taking on values 0, 1, ..., M with probabilities $q_0 > 0$, $q_1 > 0$, ..., $q_M > 0$,

$$q_0 + q_1 + \ldots + q_M = 1.$$

Denote by Y_n the stocks in the evening of day n. Then we have a recurrence

$$Y_n = \min\{M, Y_{n-1} + X_n\} - \min\{1, Y_{n-1} + X_n\}, \quad n = 1, 2, \ldots. \qquad (6.7)$$

Then $0 \leqslant Y_n \leqslant M-1$. Transitional probabilities should be calculated taking into account recurrent equation (6.7). We have:

$$\begin{aligned}
&\mathbf{P}(Y_n = 0 \mid Y_{n-1} = 0)\\
&\qquad = \mathbf{P}(\min\{M, Y_{n-1} + X_n\} - \min\{1, Y_{n-1} + X_n\} = 0 \mid Y_n = 0)\\
&\qquad = \mathbf{P}(\min\{M, X_n\} - \min\{1, X_n\} = 0 \mid Y_n = 0)\\
&\qquad = \mathbf{P}(X_n - \min\{1, X_n\} = 0)\\
&\qquad = \mathbf{P}(X_n \leqslant 1)\\
&\qquad = q_0 + q_1,\\
&\mathbf{P}(Y_n = 0 \mid Y_{n-1} = 0)\\
&\qquad = \mathbf{P}(X_n - \min\{1, X_n\} = 1)\\
&\qquad = \mathbf{P}(X_n = 2) = q_2,
\end{aligned}$$

and so on. In result, the transition probability matrix is

$$P = \begin{pmatrix} q_0 + q_1 & q_2 & \dots & q_M & 0 \\ q_0 & q_1 & \dots & q_{M-1} & q_M \\ 0 & q_0 & \dots & q_{M-2} & q_{M-1} + q_M \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & q_0 & q_1 + \dots + q_M \end{pmatrix}.$$

The reader is advised to verify that P^M has only strictly positive entries. So, Theorem 21 can be applied again. The stationary probabilities π_0, π_1, $\dots$, π_{M-1} are to be found from a system of linear equations

$$\begin{aligned}
\pi_0 &= \pi_0(q_0 + q_1) + \pi_1 q_0,\\
\pi_1 &= \pi_0 q_2 + \pi_1 q_1 + \pi_2 q_0,\\
\dots&,\\
\pi_{M-1} &= \pi_1 q_M + \pi_2(q_{M-1} + q_M) + \dots + \pi_{M-1} q_1,\\
1 &= \pi_0 + \pi_1 + \dots + q_{M-1}.
\end{aligned}$$

Solving the first $M-1$ equations successively for π_1, π_2 $\dots$, π_{M-1} and keeping π_0 as a parameter, we get

$$\begin{aligned}
\pi_1 &= \frac{1 - q_0 - q_1}{q_0}\,\pi_0,\\
\pi_2 &= \frac{(1 - q_1)(1 - q_0 - q_1) - q_0 q_2}{q_0^2}\pi_0,\\
&\dots
\end{aligned}$$

Then substituting these formulae into the normalization condition

$$1 = \pi_0 + \pi_1 + \ldots + q_{M-1}$$

we get the formula for π_0.

Now that we have the stationary probability distribution of Y_1, Y_2, $\ldots$, we are able to calculate the steady-state mathematical expectation

$$m_{\text{steady}} = \pi_1 + 2\pi_2 + 3\pi_3 + \ldots + (M-1) \cdot \pi_{M-1}$$

and the steady-state variance

$$\sigma^2_{\text{steady}} = (0 - m_{\text{steady}})^2 \pi_0 + (1 - m_{\text{steady}})^2 \pi_1 + \ldots + (M - 1 - m_{\text{steady}})^2 \cdot \pi_{M-1}$$

for remaining stocks. If a holding cost of one item is known, then one can find a confidence interval for steady-state daily expenses, etc. ■□■

Lecture 7 Statistical estimation and hypothesis testing

Statistics arose in Modern age in 17th century as a study of countries' state, especially from the point of view of economics and demographics. Introduction of advanced mathematical methods in it took place in the course of the 19th century and it gave birth to a new branch of mathematical sciences, the *mathematical statistics.* Mathematical statistics studies generally those methods and techniques used in *applied statistics*, while applied statistics uses those methods and techniques to analyse data. Mathematical statistics employs heavily methods of probability theory, it is one of the largest consumers of probability theory together with applications of stochastic processes to oprational research.

One of frequent tasks is choosing a probability model for the observed data. It can also be called *smoothing* of observed frequency distributions, or *fitting* a probability distribution to data.

Suppose we have a discrete variable X with observed values

$$x_1, \quad x_2, \quad \dots, \quad x_n \tag{7.1}$$

not necessarily different. Let us denote by u_1, u_2, ..., u_k different values among them. Denote by n_j the number of occurrences of the value u_j, $j = 1, 2, \dots, k$. Then a table with the values and their relative frequencies

$$\begin{array}{|c|c|c|c|} \hline u_1 & u_2 & \dots & u_k \\ \hline \frac{n_1}{n} & \frac{n_2}{n} & \dots & \frac{n_k}{n} \\ \hline \end{array} \tag{7.2}$$

is called the sample frequency distribution. Let us describe here *the method of moments* (method of analogy). We want to replace the observed relative frequencies n_j/n, $j = 1, 2, \dots, k$, by numbers p_j, $j = 1, 2, \dots, k$ which can be computed from a 'nice' formula and make a discrete probability distribution. Assume the formula includes a set of parameters besides j, call them θ_1, θ_2, ..., θ_r. Let us choose these parameters so that certain numerical characteristics of the probability distribution p_1, p_2, ... and those of the frequency distribution n_j/n, $j = 1, 2, \dots, k$ are same. Indeed, the mathematical expectation $m = m(\theta_1, \theta_2, \ldots, \theta_r)$, the variance $\sigma^2 = \sigma^2(\theta_1, \theta_2, \ldots, \theta_r)$ etc. become functions of the parameters. A mathematical expectation from (7.2) is

$$u_1 \frac{n_1}{n} + u_2 \frac{n_2}{n} + \ldots + u_k \frac{n_k}{n} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \overline{x},$$

i.e. the average of the observed values (7.1). It is called the sample mean. A variance from (7.2) is

$$\begin{aligned}(u_1 - \overline{x})^2 \frac{n_1}{n} + (u_2 - \overline{x})^2 \frac{n_2}{n} + \ldots + (u_k - \overline{x})^2 \frac{n_k}{n} \\ = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \ldots + (x_n - \overline{x})^2}{n} = s^2,\end{aligned}$$

So, we have equations

$$\begin{aligned}\overline{x} &= m(\theta_1, \theta_2, \ldots, \theta_r), \\ s^2 &= \sigma^2(\theta_1, \theta_2, \ldots, \theta_r), \\ &\ldots\end{aligned}$$

Instead of a variance one may use moments of the probability distribution and *sample moments*

$$\overline{x^k} = \frac{x_1^k + x_2^k + \ldots + x_n^k}{n}.$$

Example 44: Parameter estimates for some discrete probability distributions. For a binomial probability distribution with parameters N and p and known value m the unknown parameter is p, the mathematical expectation is $m = m(p) = Np$, the variance is $\sigma^2(p) = Mp(1-p)$. Parameter p can be found then either from an equation

$$Np = \overline{x}$$

or from an equation

$$Np(1-p) = s^2.$$

The first equation has the solution $p = \overline{x}/N$, but the second equation has two solutions

$$p = \frac{1}{2} - \frac{\sqrt{1 - 4s^2/N}}{2N} \qquad \text{and} \qquad p = \frac{1}{2} + \frac{\sqrt{1 - 4s^2/N}}{2N}.$$

We can't choose any without additional information on the magnitude of p.

For the Poisson probability distribution with parameter λ the mathematical expectation is $m(\lambda) = \lambda$ and the variance is $\sigma^2(\lambda) = \lambda$. So, the equation could be either

$$\lambda = \overline{x}$$

or

$$\lambda = s^2.$$

The solution is obvious. ■□■

If X is a continuous variable then the sample values (7.1) are likely to be different from each other. As we have explained in the first lecture, frequencies of hitting intervals are of interest for continuous variables. Denoting by $n(u, v)$ the number of sample values in the interval (u, v), $-\infty < u < v < \infty$, we wish to smooth frequencies $n(u, v)/n$ by means of an integral

$$\int_u^v p(x)\,dx$$

of some probability density. Somehow we identify the family of densities (i.e., normal, exponential, etc.) and then choose parameters of the probability density such that selected theoretical and sample moments are equal.

Example 45: Smoothing the Groom's age data. We wish to fit a log-normal probability density from Example 40 to the Groom's age data in Example 1. The mathematical expectation of a log-normal probability density

$$p(u) = \frac{1}{u\sigma\sqrt{2\pi}} e^{-(\ln u - m)^2/2\sigma^2}, \qquad u > 0,$$

is

$$\int_0^\infty u\,p(u)\,du = e^{m+\sigma^2/2},$$

and the variance is

$$\int_0^\infty u^2\,p(u)\,du - e^{2m+\sigma^2} = (e^{\sigma^2} - 1)e^{\sigma^2+2m}.$$

As we have already found in Example 31, the sample mean for year 2009 is 32.009 while the sample variance is 116.559. The moments equations are

$$e^{m+\sigma^2/2} = 32.009, \qquad (e^{\sigma^2} - 1)e^{\sigma^2+2m} = 116.559,$$

and the parameter estimates are $a = 3.412$ and $\sigma = 0.328$. The smoothed frequencies are shown in Table 7.1.

Table 7.1. Original and smoothed relative frequencies in intervals for the groom's ages in Costa-Rica in 2009 according to UN data

Year / Age	15–19	20–24	25–29	30–34	35–39
Original	0.0282	0.2061	0.2837	0.1865	0.1022
Smoothed	0.0862	0.1758	0.2089	0.1821	0.1317
Year / Age	40–44	45–49	50–54	55–59	>60
Original	0.0649	0.0468	0.0298	0.0202	0.0315
Smoothed	0.0849	0.0507	0.0289	0.0160	0.0348

Let us stress here that the fitted density here is just a tool to represent observed frequency distribution in a compact form. One shouldn't expect that this probability density describes the probability distribution for every single person. ■□■

The second most popular approach to fit a probability distribution to experimental data is the *maximum likelihood* approach. Assume that the sample values (7.1) were obtained in execution of an experiment n times in same conditions. We can think of the as realizations of n independent random variables X_1, X_2, ..., X_n with probability distribution identical to X. We will introduce a *likelihood function*. If X is a discrete random variable with possible values u_1, u_2, ..., and corresponding probabilities $\mathbf{P}(X = u_j) = p(u_j; \theta_1, \theta_2, \ldots, \theta_r)$ depend also on unknown parameters θ_1, θ_2, ..., θ_r, then

$$\begin{aligned} L(x_1, x_2, \ldots, x_n; \theta_1, \theta_2, \ldots, \theta_r) &= p(x_1; \theta_1, \theta_2, \ldots, \theta_r) \\ &\times p(x_2; \theta_1, \theta_2, \ldots, \theta_r) \\ &\times \ldots \times p(x_n; \theta_1, \theta_2, \ldots, \theta_r). \end{aligned}$$

For fixed values of the parameters θ_1, θ_2, ..., θ_r the likelihood function L is simply the probability to observe equalities $X_1 = x_1$, $X_2 = x_2$, ..., $X_n = x_n$, i.e. the probability to observe exactly the observed sequence of sampled values.

If X is a continuous random variable and $p(u; \theta_1, \theta_2, \ldots, \theta_r)$ is its probability density, the probability that the X_1 lies between x_1 and $x_1 + \Delta_1$, the X_2 lies between x_2 and $x_2 + \Delta_2$, and so on, until, finally, X_n lies between x_n and $x_n + \Delta_n$, is approximately

$$p(x_1; \theta_1, \theta_2, \ldots, \theta_r)\Delta_1 p(x_2; \theta_1, \theta_2, \ldots, \theta_r)\Delta_2 \times \cdots \times p(x_n; \theta_1, \theta_2, \ldots, \theta_r)\Delta_n$$

Here Δ_1, Δ_2, $\ldots$, Δ_r are small constants independent of x_1, x_2, $\ldots$, x_n. We may consider again

$$\begin{aligned} L(x_1, x_2, \ldots, x_n; \theta_1, \theta_2, \ldots, \theta_r) = {} & p(x_1; \theta_1, \theta_2, \ldots, \theta_r) \\ & \times f(x_2; \theta_1, \theta_2, \ldots, \theta_r) \\ & \times \ldots \times p(x_n; \theta_1, \theta_2, \ldots, \theta_r) \end{aligned}$$

as an estimate for likelihood of the sample (7.1).

Assigning different values to the parameters we get an idea of how probable it is to observe the sampled values (7.1) for these parameter values. Intuitively, among several events an event with the highest probability is expected to occur in one trial. The basic assumption of the method of maximum likelihood is that the sample (7.1) has been observed because of its higher probability. From this assumption we can backtrack the corresponding values of the parameters.

Definition 15. *Functions* $\hat{\theta}_1(x_1, \ldots, x_n)$, $\hat{\theta}_2(x_1, \ldots, x_n)$, $\ldots$, $\hat{\theta}_r(x_1, \ldots, x_n)$ *which maximize the likelihood function* $L(x_1, x_2, \ldots, x_n; \theta_1, \theta_2, \ldots, \theta_r)$ *are called* maximum likelihood estimators *(ML-estimator) for parameters* θ_1, θ_2 $\ldots$, θ_r.

Example 46: Maximum likelihood estimator for a probability of an event.

Let event A occur in one trial with probability p. In n trials we obtain a sequence of observations and non-observations of A. Suppose, A occurred k times. The likelihood function is

$$L(k; p) = p^k (1-p)^{n-k}.$$

To maximize it with respect to p, we can use derivative. It's more convenient to transform the likelihood function a bit: a logarithmic function is an increasing function and it maps products into sums. So,

$$\begin{aligned} \ln L(x; p) &= k \ln p + (n-k) \ln(1-p), \\ (\ln L)' &= \frac{k}{p} - \frac{n-k}{1-p} = 0, \\ (\ln L)'' &= \frac{k}{p^2} - \frac{n-k}{(1-p)^2} < 0, \end{aligned}$$

finally,

$$\hat{p} = \frac{k}{n}\,.$$

The maximum likelihood estimator for a probability is the relative frequency. ■□■

Example 47: ML-estimators for parameters of normal probability distribution. For a normal probability distribution with parameters a, σ one has

$$\begin{aligned}
L(x_1, x_2, \ldots, x_n; a, \sigma) &= \prod_{i=1}^{n} \left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-a)^2}{2\sigma^2}} \right), \\
\ln L(x_1, x_2, \ldots, x_n; a, \sigma) &= -n \ln \sigma - \frac{n}{2}\ln(2\pi) - \frac{1}{2\sigma^2}\Big\{\sum_{i=1}^{n}(x_i - a)^2\Big\}, \\
\frac{\partial \ln L}{\partial a}(x_1, x_2, \ldots, x_n; a, \sigma) &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - a) = 0, \\
\frac{\partial \ln L}{\partial \sigma}(x_1, x_2, \ldots, x_n; a, \sigma) &= -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i - a)^2 = 0.
\end{aligned}$$

Then $\hat{a}(x_1, x_2, \ldots, x_n) = \overline{x}$ and $\hat{\sigma} = s$ where s^2 is the sample variance. So, some estimators by the method of moments can also be the maximum likelihood estimators.

Note that if a is known and only σ is to be found, we have only one equation

$$\frac{\partial \ln L}{\partial \sigma}(x_1, x_2, \ldots, x_n; \sigma) = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i - a)^2 = 0$$

and its solution is

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - a)^2}\,.$$

■□■

Example 48: Maximum likelihood estimation for groom's age data. Similar calculations as above demonstrate that the maximum likelihood estimators for the parameters of log-likelihood probability distribution are

$$\hat{m}(x_1, x_2, \ldots, x_n) = \frac{1}{n}\sum_{i=1}^{n} \ln x_i,$$

$$\hat{s}(x_1, x_2, \ldots, x_n) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\ln x_i - \hat{m})^2}\,.$$

Precisely, $\hat{m} = 3.421$ and $\hat{\sigma} = 0.300$. ■□■

Example 49: Maximum likelihood estimators for parameters of a uniform distribution. Probability density of a uniform distribution in an interval (a, b) is

$$p(u; a, b) = \begin{cases} 0 & \text{when } u < a \text{ or } u > b; \\ \frac{1}{b-a} & \text{when } a \leqslant u \leqslant b. \end{cases}$$

Carefully multiplying densities we get

$$L(x_1, x_2, \ldots, x_n; a, b) = \begin{cases} 0 & \text{when } a > x_{(1)} \text{ or } b < x_{(n)}; \\ \frac{1}{(b-a)^n} & \text{when } a \leqslant x_{(1)},\ x_{(n)} \leqslant b \end{cases}$$

where $x_{(1)} = \min\{x_1, x_2, \ldots, x_n\}$, $x_{(n)} = \max\{x_1, x_2, \ldots, x_n\}$. We see that $L(x_1, x_2, \ldots, x_n; a, b)$ is a decreasing function of interval length $b - a$. But when $b - a$ is too small, $L(x_1, x_2, \ldots, x_n; a, b) = 0$. The smallest feasible interval length is $x_{(n)} - x_{(1)}$ when $a = x_{(1)}$ and $b = x_{(n)}$. ■□■

It is instructive to study the estimators as functions of the sample. In the examples above, different samples deliver different values for the estimators. Theoretically, these values can lie far from the unknown true values of the parameters θ_1, $\theta_2 <$ …, θ_r. In mathematical statistics, estimators are studies taking into account the random nature of the sample. If the sample (7.1) has been observed in independent repetitions of an experiment, independent random variables X_1, X_2, …, X_n should be substituted into an estimator formula $\theta(x_1, x_2, \ldots, x_n)$ in place of the sample values x_1, x_2, …, x_n. Thus a new random variable $\hat{\theta}(X_1, X_2, \ldots, X_n)$ appears. We will review some properties of estimators which are often respected.

Definition 16. *An estimator* $\hat{\theta}(x_1, x_2, \ldots, x_n)$ *is called* unbiased *when*

$$\mathbf{M}\hat{\theta}(X_1, X_2, \ldots, X_n) = \theta.$$

Otherwise it's called biased. *The mathematical expectation should be taken when θ is the true value of the parameter.*

Example 50: Sample moments are unbiased. Assuming a mathematical expectation $\mathbf{M}X^k$ exists,

$$\overline{X^k} = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

is an unbiased estimator for $\mathbf{M}X^k$. Indeed,

$$\begin{aligned}\mathbf{M}\overline{X^k} &= \mathbf{M}\left(\frac{1}{n}(X_1^2 + X_2^2 + \ldots + X_n^2)\right) \\ &= \frac{1}{n}(\mathbf{M}X_1^k + \mathbf{M}X_2^k + \ldots + \mathbf{M}X_n^k) = \frac{1}{n}n\mathbf{M}X = \mathbf{M}X^k.\end{aligned}$$

On the contrary, the sample variance

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

is biased (see p. 59). But it follows from computations on p. 59 that

$$S_0^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

is an unbiased estimator for the variance $\sigma^2 = \mathbf{var}\, X$. ■□■

Definition 17. *Estimator $\hat{\theta}(x_1, x_2, \ldots, x_n)$ is called* consistent *if for any $\varepsilon > 0$*

$$\lim_{n\to\infty} \mathbf{P}(|\hat{\theta}(X_1, \ldots, X_n) - \theta| > \varepsilon) = 0.$$

In other words, an estimator is consistent if it converges in probability to a true value of the parameter (see p. 59).

Example 51: Some consistent estimators. We know from Lecture 4 that sample moments are consistent estimators for the theoretical moments. We also know that sample quantiles are consistent estimators for corresponding theoretical quantities (called just *quantiles*). Under certain conditions the maximum likelihood estimates are consistent. ■□■

Definition 18. *Let $\hat{\theta}_1(x_1, x_2, \ldots, x_n)$ and $\hat{\theta}_2(x_1, x_2, \ldots, x_n)$ be two estimators for parameter θ. The estimator $\hat{\theta}_1$ is said to be* more efficient *than $\hat{\theta}_2$ if*

$$\mathbf{M}(\hat{\theta}_1 - \theta)^2 < \mathbf{M}(\hat{\theta}_2 - \theta)^2.$$

When both $\hat{\theta}_1$ and $\hat{\theta}_2$ are unbiased the condition is

$$\mathbf{var}\,\hat{\theta}_1 < \mathbf{var}\,\hat{\theta}_2.$$

The meaning of this definition is that a more efficient estimator has smaller error on the average.

Example 52: Best linear estimator for an expected value.

Let X be a random variable with finite mathematical expectation $m = \mathbf{M}X$. Then for any positive constants a_1, a_2, $\ldots$, a_n such that

$$a_1 + a_2 + \ldots + a_n = 1$$

a linear combination

$$a_1X_1 + a_2X_2 + \ldots + a_nX_n$$

is an unbiased estimator for $\mathbf{M}X$. Then, what constants should we choose best? Let us introduce an objective function

$$g(a_1, a_2, \ldots, a_n) = \mathbf{M}(a_1X_1 + a_2X_2 + \ldots + a_nX_n - m)^2,$$

then we have

$$\begin{aligned} g(a_1, a_2, \ldots, a_n) &= \mathbf{M}(a_1(X_1 - m) + a_2(X_2 - m) + \ldots + a_n(X_n - m))^2 \\ &= \sum_{i=1}^{n} a_i^2 \mathbf{M}(X_i - m)^2 + 2\sum_{i<j} a_ia_j \mathbf{M}(X_i - m)(X_j - m) = \\ &= (a_1^2 + a_2^2 + \ldots + a_n^2)\mathbf{var}\,X_1. \end{aligned}$$

We have to minimize a multivariate function $g(a_1, \ldots, a_n)$, subject to a linear constraint $a_1 + \ldots + a_n = 1$. Introducing a Lagrange multiplier λ we write

$$\mathcal{L} = (a_1^2 + \ldots + a_n^2) - \lambda(a_1 + \ldots + a_n - 1)$$

The necessary condition of extremum is $\partial\mathcal{L}/\partial a_i = 0$ for $1 \le i \le n$. We get

$$\frac{\partial\mathcal{L}}{\partial a_i} = 2a_i - \lambda = 0,$$

whence $a_1 = a_2 = \ldots = a_m = \lambda/2$. Substituting that into the constraint we have

$$\lambda = 2/n$$

and finally $a_i = 1/n$, $a_1X_1 + \ldots + a_nX_n = \overline{X}$. ■□■

Quality of approximation of experimental frequency distribution by a probability distribution can be characterized numerically. Goodness-of-fit tests were designed for this purpose. We will the Pearson's[1] chi-square test. The simplest version of the test was presented in Example 17.

Let X be a random variable and x_1, x_2, $\ldots$, x_n is a sample of its values. The problem is to see if a selected probability distribution could have generated these values. Put $z_0 = -\infty$, $z_k = \infty$. Partition the real line into intervals

$$(-\infty, z_1), [z_1, z_2), [z_3, z_4), \ldots, [z_{k-1}, \infty)$$

and compute numbers of observations in each interval,

$$n_1, n_2, n_3, \ldots, n_k.$$

In practice, the number k of intervals and the dissection points z_1, z_2, $\ldots$, z_{k-1} should be such that each interval holds at least five observations (this magic constant has no theoretical background and is a suggestion of practitioners). Now compute probabilities for each interval,

$$p_i = \mathbf{P}(z_{i-1} \leqslant X < z_i)$$

and *expected frequencies*

$$e_i = np_i.$$

If the selected probability distribution fits well with the data, n_i, should be close to the expected frequencies, e_i. In fact, the actual number of observations in the interval $[z_{i-1}, z_i)$ is random and it has the binomial distribution with parameters n, p_i. Then $np_i = e_i$ is the mathematical expectation of the number of such observations.

The goodness of fit is measured with statistic

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - e_i)^2}{e_i} = n \sum_{i=1}^{k} \frac{\left(p_i - \frac{n_i}{n}\right)^2}{p_i}.$$

Assuming the hypothesis on the probability distribution of X is true, the probability distribution of χ^2 statistic tends to the χ^2 distribution with

[1] Pearson Karl (1857–1936) was English mathematician and statistician.

$k-1$ degrees of freedom. Since small values of χ^2 agree with the hypothesized probability distribution of X, we will reject the hypothesis when χ^2 is greater than some limit R. This decision rule introduces a possibility of an error when we reject a true hypothesis. This type of errors is called *type I*. The probability of a type I error is called a *significance level* of the test (of the decision rule). Since under the hypothesis being tested the χ^2 statistic has approximately χ^2 distribution with parameter $k-1$, the corresponding probability can be expressed by an integral over the χ^2 probability density,

$$\int\limits_R^\infty \frac{u^{(k-1)2-1}e^{-u/2}}{2^{(k-1)/2}\Gamma((k-1)/2)}\,du.$$

The integral tends to zero as R grow to infinity. So, choosing a value R sufficiently large, we can make the integral together with the type I error less or equal a previously selected significance level α.

If the hypothesized probability distribution is known up to a number l of parameters and the parameters have been estimated my method of maximum likelihood then the number of degrees of freedom should be reduced by l.

Example 53: Testing goodness-of-fit for groom's age data. In Example 45 the observed frequencies were smoothed by a log-normal probability density with parameters $m = 3.412$ and $\sigma = 0.3$. The intervals, the observed and the predicted frequencies are shown in Table 7.2.

Table 7.2. Observed and predicted frequencies according to a log-normal probability distribution

Year / Age	$(-\infty, 19)$	$[20, 25)$	$[25, 30)$	$[30, 35)$	$[35, 40)$
observed	0.0282	0.2061	0.2837	0.1865	0.1022
predicted	0.0782	0.1721	0.2234	0.1992	0.1412
Year / Age	$[40, 45)$	$[45, 50)$	$[50, 55)$	$[55, 60)$	$[60, \infty)$
Original	0.0649	0.0468	0.0298	0.0202	0.0315
Smoothed	0.0866	0.0485	0.0255	0.0129	0.0124

The total number of observations is 23 221 and the χ^2 statistic is

$$23221\cdot\Bigg(\frac{(0.0282-0.0782)^2}{0.0782}+\frac{(0.2061-0.1720)^2}{0.1720}+\frac{(0.2837-0.2234)^2}{0.22324}$$

$$+\frac{(0.1865-0.1992)^2}{0.1992}+\frac{(0.1022-0.1412)^2}{0.1412}+\frac{(0.0649-0.0866)^2}{0.0866}$$

$$+\frac{(0.0468-0.0485)^2}{0.0485}+\frac{(0.0298-0.0255)^2}{0.0255}+\frac{(0.0202-0.0129)^2}{0.0129}$$
$$\left.+\frac{(0.0315-0.0124)^2}{0.0124}\right)=2472.$$

This value is too large. The number of degrees of freedom is $10-2-1=7$ and with probability 0.99 a random variable with the χ^2-distribution with 7 degrees of freedom lies in the interval $(0, 18.475)$, i.e. $R = 18.475$ for the significance level $\alpha = 0.01$. The log-normal hypothesis should be rejected.

In such a situation a statistician tries another probability distribution. Let us try and fit a shifted gamma-distribution with probability density

$$p(u)=\frac{(u-\tau)^{a-1}}{b^a\Gamma(b)}e^{-(u-\tau)/b}, \qquad u>\tau$$

where τ is the shift parameter, a is called a shape parameter, and b is called a scale parameter. We let $\tau = 15$ which corresponds to the smallest observed value of 15 years. Unfortunately, the maximum likelihood estimators are solutions of a transcendent system of equations has no explicit solution, so the parameters a and b should be found by a numerical procedure. We get, after all, $a = 2.835$ and $b = 0.1658$. The predicted frequencies are shown in Table 7.3. The χ^2 statistic is 1380.16 which is still too large but is lower than that for a log-normal probability distribution.

Table 7.3. Observed and predicted frequencies according to a shifted gamma probability distribution

Year / Age	$(-\infty, 19)$	$[20, 25)$	$[25, 30)$	$[30, 35)$	$[35, 40)$
observed	0.0282	0.2061	0.2837	0.1865	0.1022
predicted	0.0660	0.2001	0.2274	0.1856	0.1294
Year / Age	$[40, 45)$	$[45, 50)$	$[50, 55)$	$[55, 60)$	$[60, \infty)$
Original	0.0649	0.0468	0.0298	0.0202	0.0315
Smoothed	0.0819	0.0487	0.0277	0.0153	0.0169

Sometimes a feasible family of probability distributions is known from general considerations and a statistician is only unsure about the value of a single parameter θ. We will consider the simplest case when to choose between two possibilities, θ_0 and θ_1. Let us call the assumption $\theta = \theta_0$ the *null hypothesis* and the assumption $\theta = \theta_1$ the *alternative hypothesis*. A decision rule observes the sample (7.1) and decides to accept H_0 or reject H_0

in favor of H_1. A general method to construct a decision rule by considering the *likelihood ratio*

$$\frac{L(x_1, x_2, \dots, x_n; \theta_1)}{L(x_1, x_2, \dots, x_n; \theta_0)}.$$

In mathematical statistics it is proven that the optimal decision rule rejects H_0 when the likelihood ratio exceeds a level C_α. The level should be chosen to guarantee the significance level α. Besides type I errors, other error is possible. Assume H_1 is true but the observed sample is such that the decision rule prescribes to accept H_0. This is a type II error. Its probability is often denoted by β.

Example 54: Testing hypothesis on the mathematical expectation of normal probability distribution. Let X have a normal probability distribution with unknown mathematical expectation a and known variance σ^2. Let the possible values for$a$ be a_0 and a_1. Then the likelihood ratio is

$$\begin{aligned}\frac{L(x_1, x_2, \dots, x_n; \theta_1)}{L(x_1, x_2, \dots, x_n; \theta_0)} &= \frac{\prod_{j=1}^n \dfrac{1}{\sigma\sqrt{2\pi}} e^{-(x_i-a_1)^2/2\sigma^2}}{\prod_{j=1}^n \dfrac{1}{\sigma\sqrt{2\pi}} e^{-(x_i-a_0)^2/2\sigma^2}} \\ &= \exp\Big\{\frac{1}{2\sigma^2}\sum_{i=1}^n((x_i-a_0)^2-(x_i-a_1)^2)\Big\} \\ &= \exp\Big\{\frac{n(a_1-a_0)}{\sigma^2}\overline{x}+\frac{n}{2\sigma^2}(a_0^2-a_1^2)\Big\}.\end{aligned}$$

Let us assume additionally that $a_1 > a_0$ (the opposite case $a_1 < a_0$ should be studied by the reader on his own). Then the likelihood ratio is an increasing function of $\overline{x}$. The inequality

$$\frac{L(x_1, x_2, \dots, x_n; \theta_1)}{L(x_1, x_2, \dots, x_n; \theta_0)} > C_\alpha$$

is equivalent to

$$\overline{x} > \widetilde{C}_\alpha$$

where $\widetilde{C}_\alpha$ should be chosen to guarantee the significance level α. We know that the sample mean $\overline{X}$ has the normal probability distribution with mathematical expectation a_0 (under H_0) and variance σ^2/n (see Example 23). Hence, $\widetilde{C}_\alpha$ is the solution of the equation

$$1-\Phi\Big(\frac{\widetilde{C}_\alpha-a_0}{\sigma}\sqrt{n}\Big)=\alpha,$$

i.e. $\widetilde{C}_\alpha = a_0 + \sigma z_{1-\alpha}/\sqrt{n}$ where $z_{1-\alpha}$ is the solution of $\Phi(z) = \alpha$ (determined from Table 3). Under H_1 the sample mean $\overline{X}$ has the normal probability distribution with mathematical expectation a_1 and variance σ^2/n and the probability of type II error is

$$\beta = \Phi\Big(\frac{\widetilde{C}_\alpha - a_1}{\sigma}\sqrt{n}\Big) = \Phi\Big(\frac{a_0 - a_1}{\sigma}\sqrt{n} + z_{1-\alpha}\Big) < \Phi(z_{1-\alpha}) = 1 - \alpha.$$

So, $\beta + \alpha \leqslant 1$. If we want to guarantee the type II error probability β we can do that by choice of n the sample size from the formula above. ■□■

Lecture 8 Statistical methods to process experimental data

In this lecture we will review several important statistical procedures to study functional and statistical dependence of variables.

First we will study a linear regression model. Suppose we study a production process where one input variable X affects one output variable Y. Suppose we have a sample

$$(x_1, y_1), \quad (x_2, y_2), \quad \ldots, \quad (x_n, y_n) \tag{8.1}$$

where a value of X was set and a response value of Y was measured, keeping other conditions the same. Plotting the points (8.1) we may have a guess that

$$y_j = b_0 + b_1 x_j + e_j$$

where b_0 (the intersect) and b_1 (the slope) are parameters of a linear function, and e_j is an error term due to measurement errors and influence of other uncontrolled factors. The proposed model is then

$$Y_j = b_0 + b_1 x_j + \varepsilon_j, \qquad j = 1, 2, \ldots, n,$$

where Y_j is a random variable representing the output in j-th measurement above, and ε_j is the random error term. Following Gauss, we assume that

1. The mean error is zero, $\mathbf{M}\varepsilon_j = 0$, $j = 1,\ 2,\ \ldots,\ n$;
2. The errors are uncorrelated and have same variances, $\mathbf{var}\,\varepsilon_j = \sigma^2$ and $\mathbf{cov}\,(\varepsilon_i, \varepsilon_j) = 0$, $1 \leqslant i < j \leqslant n$;
3. $x_1,\ x_2,\ \ldots,\ x_n$ are nonrandom;
4. ε_j, $j = 1,\ 2,\ \ldots,\ n$ have normal probability distribution.

Let us find maximum likelihood estimators for the intercept b_0, the slope b_1, and the standard deviation of errors σ. Under Gauss's assumptions, Y_j has a normal probability distribution with the mathematical expectation $b_0 + b_1 x_j$ and variance σ^2. The likelihood function is

$$L(b_0, b_1, \sigma) = \prod_{j=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-(y_j - b_0 - b_1 x_j)^2/2\sigma^2}.$$

and the logarithm of that is

$$\ln L(b_0, b_1, \sigma) = -\frac{n}{2} \ln 2\pi\sigma^2 - \frac{1}{2\sigma^2} \sum_{j=1}^{n} (y_j - b_0 - b_1 x_j)^2.$$

By taking derivatives of $\ln L(b_0, b_1, \sigma)$ with respect to b_0, b_1, and σ^2 we have equations

$$\begin{aligned}
\frac{\partial \ln L}{\partial b_0} &= \frac{1}{\sigma^2} \sum_{j=1}^{n} (y_j - b_0 - b_1 x_j) = 0, \\
\frac{\partial \ln L}{\partial b_1} &= \frac{1}{\sigma^2} \sum_{j=1}^{n} x_j (y_j - b_0 - b_1 x_j) = 0, \\
\frac{\partial \ln L}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{j=1}^{n} (y_j - b_0 - b_1 x_j)^2 = 0.
\end{aligned}$$

The solution is

$$\begin{aligned}
\hat{b}_1 &= \frac{n \sum\limits_{j=1}^{n} x_j y_j - \Big(\sum\limits_{j=1}^{n} x_j \Big) \Big(\sum\limits_{j=1}^{n} y_j \Big)}{n \sum\limits_{j=1}^{n} x_j^2 - \Big(\sum\limits_{j=1}^{n} x_j \Big)^2} = \frac{\frac{1}{n} \sum\limits_{j=1}^{n} (x_j - \overline{x}) y_j}{s_{\mathrm{x}}^2}, \\
\hat{b}_0 &= \overline{y} - \hat{b}_1 \overline{x} = \frac{1}{n s_{\mathrm{x}}^2} \sum_{j=1}^{n} y_j (s_{\mathrm{x}}^2 - (x_j - \overline{x})), \\
\hat{\sigma}^2 &= \frac{1}{n} \sum_{j=1}^{n} (y_j - \hat{b}_0 - \hat{b}_1 x_j)^2,
\end{aligned}$$

where

$$s_{\mathrm{x}}^2 = \frac{1}{n} \sum_{j=1}^{n} x_j^2 - \Big(\frac{1}{n} \sum_{i=1}^{n} x_j \Big)^2, \qquad s_{\mathrm{x}} = \sqrt{s_{\mathrm{x}}^2}$$

is the sample variance of X. Functions $\hat{b}_0 = \hat{b}_0(y_1, y_2, \dots, y_n)$ and $\hat{b}_1 = \hat{b}_1(y_1, y_2, \dots, y_n)$ are linear in the variables y_1, y_2, $\dots$, y_n. So, they have normal probability distributions with mathematical expectations

$$\mathbf{M}\hat{b}_1(Y_1, Y_2, \dots, Y_n) = \frac{n \sum\limits_{j=1}^{n} x_j \mathbf{M} Y_j - \Big(\sum\limits_{j=1}^{n} x_j \Big) \Big(\sum\limits_{j=1}^{n} \mathbf{M} Y_j \Big)}{n \sum\limits_{j=1}^{n} x_j^2 - \Big(\sum\limits_{j=1}^{n} x_j \Big)^2}$$

$$= \frac{n\sum\limits_{j=1}^{n} x_j(b_0+b_1x_j) - \Big(\sum\limits_{j=1}^{n} x_j\Big)\Big(\sum\limits_{j=1}^{n}(b_0+b_1x_j)\Big)}{n\sum\limits_{j=1}^{n} x_j^2 - \Big(\sum\limits_{j=1}^{n} x_j\Big)^2} = b_1,$$

$$\begin{aligned}\mathbf{M}\hat{b}_0(Y_1,Y_2,\dots,Y_n) &= \mathbf{M}\overline{Y} - \overline{x}\mathbf{M}\hat{b}_1(Y_1,Y_2,\dots,Y_n)\\ &= b_0 + b_1\overline{x} - \overline{x}b_1 = b_0,\end{aligned}$$

and variances

$$\begin{aligned}\mathbf{var}\,\hat{b}_1(Y_1,Y_2,\dots,Y_n) &= \mathbf{var}\,\frac{\sum\limits_{j=1}^{n} Y_j(x_j-\overline{x})}{ns_{\mathrm{x}}^2}\\ &= \frac{\sum\limits_{j=1}^{n}(x_j-\overline{x})^2\mathbf{var}\,Y_j}{n^2s_{\mathrm{x}}^4} = \frac{\sigma^2}{ns_{\mathrm{x}}^2},\\ \mathbf{var}\,\hat{b}_0(Y_1,Y_2,\dots,Y_n) &= \frac{1}{n^2s_{\mathrm{x}}^4}\sum_{j=1}^{n}(s_{\mathrm{x}}^2-(x_j-\overline{x}))^2\mathbf{var}\,Y_j\\ &= \frac{\sigma^2(s_{\mathrm{x}}^2+1)}{ns_{\mathrm{x}}^2}.\end{aligned}$$

It can be proven also that $n\hat{\sigma}^2(Y_1,Y_2,\dots,Y_n)/\sigma^2$ has the χ^2-distribution with $n-2$ degrees of freedom. It turns out that the random variables $\hat{b}_0(X_1,X_2,\dots,X_n)$ and $\hat{\sigma}^2(X_1,X_2,\dots,X_n)$ are independent, and so are $\hat{b}_1(X_1,X_2,\dots,X_n)$ and $\hat{\sigma}^2(X_1,X_2,\dots,X_n)$. It follows from Lecture 3 that a statistic

$$\frac{\hat{b}_1(Y_1,Y_2,\dots,Y_n)-b_1}{\hat{\sigma}(Y_1,Y_2,\dots,Y_n)}\sqrt{ns_{\mathrm{x}}^2}$$

has Student's t-distribution with $(n-2)$ degrees of freedom.

The quality of linear model for the sample (8.1) can be measured as follows. Let us consider the total sum of squares

$$Q = \sum_{j=1}^{n}(y_j-\overline{y})^2$$

of deviations of the output values from their average. Using explicit formulae for $\hat{b}_0$ and $\hat{b}_1$ we can partition the total sum of squares into a sum

$$Q_1+Q_2 = \sum_{j=1}^{n}(\hat{b}_0+\hat{b}_1x_j-\overline{y})^2 + \sum_{j=1}^{n}(y_j-\hat{b}_0-\hat{b}_1x_j)^2.$$

Here Q_1 is the *regression sum of squares*, or *explained sum of squares*, and Q_2 is the *sum of squares of residuals*, or the *residual sum of squares.* If the linear dependence on X explains the observed changes in Y, then Q_1 should be close to Q and Q_2 should be small. The *coefficient of determination* is defined as

$$R^2 = \frac{Q_1}{Q} = 1 - \frac{Q_2}{Q}.$$

The closer R^2 is to unity the better.

Under a hypothesis

$$H_0 \colon b_1 = 0$$

the ratio

$$F = \frac{(n-2)Q_1}{Q_2}$$

has an F-distribution with 1 and $n-2$ degrees of freedom. Here hypothesis H_0 means no linear dependence of Y on X. H_0 is rejected when the observed value F exceeds a table value for a selected significance level α.

After the parameters b_0 and b_1 have been estimated and H_0 has been rejected one uses the equation

$$Y = \hat{b}_0 + \hat{b}_1 X$$

to predict expected values of Y for different values of X within the studied range of X.

Example 55: Production cost of a book. Table 8.1 demonstrates the production cost y (in hundred Roubles) of one book item depending on a number x of printed copies (in thousands) over several years. The plot of the data in Fig. 8.1 demonstrates that the production cost is in inverse ratio to the number of prined copies. Firstly, we change the input variable x into $z = 1/x$. The plot of y as a function of z can be seen in Fig. 8.2.

Table 8.1. Production cost of a book item for different numbers of printed copies in thousands ($z_i = 1/x_i$)

x_i	1	2	3	5	10	20	30	50	100	200
y_i	10.15	5.52	4.08	2.85	2.11	1.62	1.41	1.30	1.21	1.15
z_i	1	0.5	0.333	0.2	0.1	0.05	0.033	0.02	0.01	0.005

It is natural to seek for a linear function

$$y = b_0 + b_1 z = b_0 + \frac{b_1}{x}\,.$$

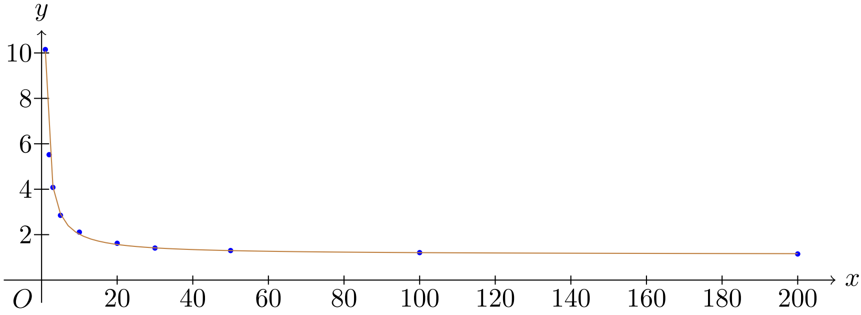

Figure 8.1. Plot of production cost of an item (blue points) and a fitted curve (brown)

Using the formulae above we obtain

$$\begin{aligned}\overline{z} &= 0.225,\\ \overline{y} &= 3.14,\\ \sum_{j=1}^{11} x_j y_j &= 15.223,\\ \hat{b}_1 &= 8.876,\\ \hat{b}_0 &= 1.119,\\ \hat{\sigma}^2 &= 0.03424.\end{aligned}$$

The fitted curve is

$$y = 1.119 + 8.876z = 1.119 + \frac{8.876}{z}\,.$$

A brown curve in Fig. 8.1 and a brown line in 8.2 show the estimated functional dependence. To compute the coefficient of determination we compute Q and Q_2. We have

$$Q = 73.207, \qquad Q_2 = 0.0274, \qquad R^2 = 0.9996.$$

The F-ratio is 21376.3 which is rather high for a random variable with F-distribution with 8 and 1 degrees of freedom. Here b_0 can be interpreted as the cost of production of one item in an infinite print run. Please pay

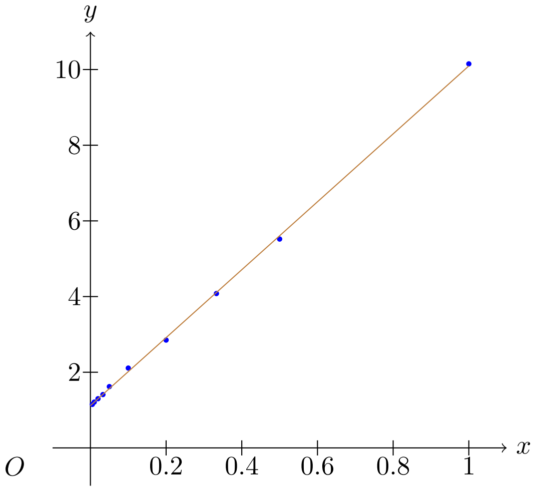

Figure 8.2. Plot of production cost of an item after change of the input variable

attention that this formula behaves badly as x is near zero. The reasons are both economic and mathematical. Producing half a book is meaningless (we try to apply the formula in wrong domain), and the regression is not a good tool for *extrapolation*. ■□■

This simple regression model with one input variable and one output variable has a natural extension to several input variables. We will consider below a particular case when the input variables take on 'binary values' 0 and 1. Such a model can be used to compare several groups of objects. Consider the following situation: we have g different technologies to produce an item and the question is if all technologies guarantee the same reliability, X, of a product. We can study items produced using each technology $i = 1$, $2, \ldots, g$, and measure the reliability. Denote by $x_{i,j}$ the reliability of the j-th item in the i-th group (i-th technology). Let the i-th group have n_i observations. Here a technology type is the factor explaining the variable X.

A model with one *factor* (i.e. a qualitative input variable) is

$$X_{i,j} = \mu_i + \varepsilon_{i,j},$$

where μ_i is a mean value explained by the value of the factor, and $\varepsilon_{i,j}$ is a random measurement error.

Classic assumptions are:

1. $\varepsilon_{i,j} = 0$, so that there's no systematic errors in observations;
2. errors $\{\varepsilon_{i,j} \colon i = 1, 2, \ldots, g; j = 1, 2, \ldots, n_i\}$ are independent;
3. the errors have the same variance, $\mathbf{var}\, \varepsilon_{i,j} = \sigma^2$;
4. the errors $\{\varepsilon_{i,j} \colon i = 1, 2, \ldots, g; j = 1, 2, \ldots, n_i\}$ have a normal distribution.

A null hypothesis is

$$H_0 \colon \mu_1 = \mu_2 = \ldots = \mu_g,$$

in other words, all groups are statistically similar, and the factor doesn't influence the variable X.

Put $n = n_1 + n_2 + \ldots + n_g$,

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{g} \sum_{j=1}^{n_i} x_{i,j}, \qquad \overline{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{i,j}.$$

Here $\overline{x}$ is the overall average, $\overline{x}_i$ is the average in the i-th group. Then we transform the total sum of squares of deviations as

$$\sum_{i=1}^{g} \sum_{j=1}^{n_i} (x_{i,j} - \overline{x})^2 = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (x_{i,j} - \overline{x}_i + \overline{x}_i - \overline{x})^2 =$$

$$= \sum_{i=1}^{g} \sum_{j=1}^{n_i} (x_{i,j} - \overline{x}_i)^2 + 2 \sum_{i=1}^{g} \sum_{j=1}^{n_i} (x_{i,j} - \overline{x}_i)(\overline{x}_i - \overline{x}) + \sum_{i=1}^{g} \sum_{j=1}^{n_i} (\overline{x}_i - \overline{x})^2.$$

One has

$$\sum_{i=1}^{g} \sum_{j=1}^{n_i} (x_{i,j} - \overline{x}_i)(\overline{x}_i - \overline{x}) = \sum_{i=1}^{g} \left((\overline{x}_i - \overline{x}) \sum_{j=1}^{n_g} (x_{i,j} - \overline{x}_i) \right) = 0.$$

Thus

$$\underbrace{\sum_{i=1}^{g} \sum_{j=1}^{n_i} (x_{i,j} - \overline{x})^2}_{Q_0} = \underbrace{\sum_{i=1}^{g} \sum_{j=1}^{n_i} (x_{i,j} - \overline{x}_i)^2}_{Q_1} + \underbrace{\sum_{i=1}^{g} n_i (\overline{x}_i - \overline{x})^2}_{Q_2}.$$

Q_0 is called *a total sum of squares*, Q_1 is called *a sum of squares within groups*, Q_2 is called *a weighted sum of squares between groups.*

When H_0 is true, it seems likely that Q_0 and Q_1 should be comparatively close while Q_2 should be significantly less that Q_1.

To test H_0 a ratio

$$F = \frac{Q_2(n-g)}{Q_1(g-1)}$$

is used. Under H_0, F has an *F-distribution* with $g-1$ and $n-g$ degrees of freedom. H_0 is rejected at significance level α, when $F > F_{\alpha;g-1,n-g}$ for suitably chosen $F_{\alpha;g-1,n-g}$ (using tables or software).

Example 56: Teaching method study. Four groups of students were taught to perform a production operations. Each group had its own study program. After completion of the learning process the students had to produce articles during one hour. Table 8.2 demonstrates the numbers of articles produces by the students. We'd like to prove that some of the teaching methods lead to higher productivity.

Table 8.2. Productivity of students in different groups

Group	$x_{i,j}$	n_i	group average, $\overline{x}_i$
1	60, 80, 75, 80, 85, 70	6	75
2	75, 66, 85, 80, 70, 80, 90	7	78
3	60, 80, 65, 60, 86, 75	6	71
4	95, 85, 100, 80	4	90
	Total:	23	77.48

The computations necessary to test the hypothesis that the teaching methods don't lead to different resulting productivities of students:

$$\overline{x}_1 = \frac{1}{6}(60+80+75+80+85+70) = 75,$$

$$\overline{x}_2 = \frac{1}{7}(75+66+85+80+70+80+90) = 78,$$

$$\overline{x}_3 = \frac{1}{6}(60+80+65+60+86+75) = 71,$$

$$\overline{x}_4 = \frac{1}{4}(95+85+100+80) = 90,$$

$$\overline{x} = \frac{1}{23}(60 + 80 + \ldots + 80) = 77.48,$$

$$Q_0 = (60 - 77.48)^2 + (80 - 77.48)^2 + \ldots + (80 - 77.48)^2 = 2414.7,$$

$$\begin{gathered} Q_1 = (60 - 75)^2 + \ldots + (70 - 75)^2 + \\ +(75 - 78)^2 + \ldots + (90 - 78)^2 + \\ +(60 - 71)^2 + \ldots + (75 - 71)^2 + \\ +(95 - 90)^2 + \ldots + (80 - 90)^2 = 1497, \end{gathered}$$

$$\begin{gathered} Q_2 = 6(75 - 77.48)^2 + 7(78 - 77.48)^2 + \\ +6(71 - 77.48)^2 + 4(90 - 77.48)^2 = 1497, \end{gathered}$$

$$F = \frac{1497 \cdot (23 - 4)}{1497 \cdot (4 - 1)} = 3.88.$$

For $\alpha = 0.05$, $F_{0.05;3,19} = 3.127$. H_0 should be rejected. ■□■

When the variables are qualitative (taking non-numerical values), one can't speak of functional dependence. Here statistical dependence comes in handy. Assume n items have been observed and two variables describing an item were recorded. Variables can be both numerical (e.g. a measurement) and categorical (e.g. color, sex, etc.) The values of the first variable are distributed in k classes (intervals), the values of the second variables are distributed in m classes. Data is represented in a *contingency table.*

	B_1	B_2	$\ldots$	B_m
A_1	$n_{1,1}$	$n_{1,2}$	$\ldots$	$n_{1,m}$
A_2	$n_{2,1}$	$n_{2,2}$	$\ldots$	$n_{2,m}$
$\vdots$	$n_{1,1}$	$n_{1,2}$	$\ldots$	$n_{1,m}$
A_k	$n_{k,1}$	$n_{k,2}$	$\ldots$	$n_{k,m}$

Here $n_{i.j}$ is the number of observations belonging to the i-th class of the first variable and at the same time to the j-th class of the second variable.

Denote by $p_{i,j} = \mathbf{P}(A_i, B_j)$ the probability for an item to belong to classes A_i and B_j at the same time, $p_i^{(1)} = \mathbf{P}(A_i)$, $p_j^{(2)} = \mathbf{P}(B_j)$ with obvious meaning. We want to test a null hypothesis

$$H_0\colon p_{i,j} = p_i^{(1)} p_j^{(2)} \text{ for all } i,\ j.$$

H_0 means that the variables are independent. The maximum likelihood estimators for the probabilities are

$$\hat{p}_{i,j} = \frac{n_{i,j}}{n},$$

$$\hat{p}_i^{(1)} = \frac{n_{i,1} + \ldots + n_{i,m}}{n} = \frac{n_i^{(1)}}{n},$$

$$\hat{p}_j^{(2)} = \frac{n_{1,j} + \ldots + n_{k,j}}{n} = \frac{n_j^{(2)}}{n}.$$

Under H_0 it is likely that

$$\hat{p}_{i,j} \approx \hat{p}_i^{(1)} \hat{p}_j^{(2)},$$

or

$$n_{i,j} \approx \frac{n_i^{(1)} n_j^{(2)}}{n}.$$

To test H_0 we will use a χ^2 statistic

$$\chi^2 = n \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{\left(n_{i,j} - \frac{n_i^{(1)} n_j^{(2)}}{n}\right)^2}{n_i^{(1)} n_j^{(2)}}.$$

Under H_0 χ^2 has approximately the χ^2-square probability distribution with $(k-1)(l-1)$ degrees of freedom as $n \to \infty$. So, we can find a confidence interval for this statistics under H_0 for any significance level α and then we will reject the null hypothesis when the χ^2 statistic falls outside this interval.

References

1. Brownlee K.A. Statistical theory and methodology in sciences and engineering. — N.Y., London, Sidney: J.Wiley and Sons, 1960.

2. Everitt B.S., Hothorn T. A handbook of statistical analysis using R. — Chapman & Hall/CRC, 2010.

3. Kemeny J.C, Snell J.L. Finite Markov chains. — Princeton, NJ: Van Nostrand, 1960.

4. Kimble G.A. How to use (and misuse) Statistics. — Englewood Cliffs, N.J.: Prentice Hall, Inc., 1978.

5. Mittelhammer R.C. Mathematical statistics for economics and buisiness. — N.Y., Heidelberg, London: Springer, 2013.

6. Shiriaev A.N. Probability. — Springer, 1996.

7. Shiryaev A.N. Essentials of stochastic finance: facts, models, theory. — World Scientific, 1999.

Table 3. Table of the function $\Phi_0(u) = \dfrac{1}{\sqrt{2\pi}} \int\limits_0^u e^{-t^2/2}\,dt$

u	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.00000	0.00399	0.00798	0.01197	0.01595	0.01994	0.02392	0.02790	0.03188	0.03586
0.1	0.03983	0.04380	0.04776	0.05172	0.05567	0.05962	0.06356	0.06749	0.07142	0.07535
0.2	0.07926	0.08317	0.08706	0.09095	0.09483	0.09871	0.10257	0.10642	0.11026	0.11409
0.3	0.11791	0.12172	0.12552	0.12930	0.13307	0.13683	0.14058	0.14431	0.14803	0.15173
0.4	0.15542	0.15910	0.16276	0.16640	0.17003	0.17364	0.17724	0.18082	0.18439	0.18793
0.5	0.19146	0.19497	0.19847	0.20194	0.20540	0.20884	0.21226	0.21566	0.21904	0.22240
0.6	0.22575	0.22907	0.23237	0.23565	0.23891	0.24215	0.24537	0.24857	0.25175	0.25490
0.7	0.25804	0.26115	0.26424	0.26730	0.27035	0.27337	0.27637	0.27935	0.28230	0.28524
0.8	0.28814	0.29103	0.29389	0.29673	0.29955	0.30234	0.30511	0.30785	0.31057	0.31327
0.9	0.31594	0.31859	0.32121	0.32381	0.32639	0.32894	0.33147	0.33398	0.33646	0.33891
1.0	0.34134	0.34375	0.34614	0.34849	0.35083	0.35314	0.35543	0.35769	0.35993	0.36214
1.1	0.36433	0.36650	0.36864	0.37076	0.37286	0.37493	0.37698	0.37900	0.38100	0.38298
1.2	0.38493	0.38686	0.38877	0.39065	0.39251	0.39435	0.39617	0.39796	0.39973	0.40147
1.3	0.40320	0.40490	0.40658	0.40824	0.40988	0.41149	0.41309	0.41466	0.41621	0.41774
1.4	0.41924	0.42073	0.42220	0.42364	0.42507	0.42647	0.42785	0.42922	0.43056	0.43189
1.5	0.43319	0.43448	0.43574	0.43699	0.43822	0.43943	0.44062	0.44179	0.44295	0.44408
1.6	0.44520	0.44630	0.44738	0.44845	0.44950	0.45053	0.45154	0.45254	0.45352	0.45449
1.7	0.45543	0.45637	0.45728	0.45818	0.45907	0.45994	0.46080	0.46164	0.46246	0.46327
1.8	0.46407	0.46485	0.46562	0.46638	0.46712	0.46784	0.46856	0.46926	0.46995	0.47062
1.9	0.47128	0.47193	0.47257	0.47320	0.47381	0.47441	0.47500	0.47558	0.47615	0.47670
2.0	0.47725	0.47778	0.47831	0.47882	0.47932	0.47982	0.48030	0.48077	0.48124	0.48169
2.1	0.48214	0.48257	0.48300	0.48341	0.48382	0.48422	0.48461	0.48500	0.48537	0.48574
2.2	0.48610	0.48645	0.48679	0.48713	0.48745	0.48778	0.48809	0.48840	0.48870	0.48899
2.3	0.48928	0.48956	0.48983	0.49010	0.49036	0.49061	0.49086	0.49111	0.49134	0.49158
2.4	0.49180	0.49202	0.49224	0.49245	0.49266	0.49286	0.49305	0.49324	0.49343	0.49361
2.5	0.49379	0.49396	0.49413	0.49430	0.49446	0.49461	0.49477	0.49492	0.49506	0.49520
2.6	0.49534	0.49547	0.49560	0.49573	0.49585	0.49598	0.49609	0.49621	0.49632	0.49643
2.7	0.49653	0.49664	0.49674	0.49683	0.49693	0.49702	0.49711	0.49720	0.49728	0.49736
2.8	0.49744	0.49752	0.49760	0.49767	0.49774	0.49781	0.49788	0.49795	0.49801	0.49807
2.9	0.49813	0.49819	0.49825	0.49831	0.49836	0.49841	0.49846	0.49851	0.49856	0.49861
3.0	0.49865	0.49869	0.49874	0.49878	0.49882	0.49886	0.49889	0.49893	0.49896	0.49900
3.1	0.49903	0.49906	0.49910	0.49913	0.49916	0.49918	0.49921	0.49924	0.49926	0.49929
3.2	0.49931	0.49934	0.49936	0.49938	0.49940	0.49942	0.49944	0.49946	0.49948	0.49950
3.3	0.49952	0.49953	0.49955	0.49957	0.49958	0.49960	0.49961	0.49962	0.49964	0.49965
3.4	0.49966	0.49968	0.49969	0.49970	0.49971	0.49972	0.49973	0.49974	0.49975	0.49976
3.5	0.49977	0.49978	0.49978	0.49979	0.49980	0.49981	0.49981	0.49982	0.49983	0.49983
3.6	0.49984	0.49985	0.49985	0.49986	0.49986	0.49987	0.49987	0.49988	0.49988	0.49989
3.7	0.49989	0.49990	0.49990	0.49990	0.49991	0.49991	0.49992	0.49992	0.49992	0.49992
3.8	0.49993	0.49993	0.49993	0.49994	0.49994	0.49994	0.49994	0.49995	0.49995	0.49995
3.9	0.49995	0.49995	0.49996	0.49996	0.49996	0.49996	0.49996	0.49996	0.49997	0.49997
4.0	0.49997	0.49997	0.49997	0.49997	0.49997	0.49997	0.49998	0.49998	0.49998	0.49998
4.1	0.49998	0.49998	0.49998	0.49998	0.49998	0.49998	0.49998	0.49998	0.49999	0.49999
4.2	0.49999	0.49999	0.49999	0.49999	0.49999	0.49999	0.49999	0.49999	0.49999	0.49999
4.3	0.49999	0.49999	0.49999	0.49999	0.49999	0.49999	0.49999	0.49999	0.49999	0.49999

ВОСЕМЬ ЛЕКЦИЙ ПО ТЕОРИИ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКОЙ СТАТИСТИКЕ

Автор:
Андрей Владимирович **Зорин**

Учебно-методическое пособие

Федеральное государственное бюджетное
образовательное учреждение
высшего профессионального образования
«Нижегородский государственный университет
им. Н.И. Лобачевского».
603950, Нижний Новгород, пр. Гагарина, 23