Филиал Федерального государственного автономного
образовательного учреждения
высшего образования "Российский государственный
университет нефти и газа
(национальный исследовательский университет)
имени И.М. Губкина" в городе Ташкенте

Отделение "Математика и информатика"

N.G. Gamkrelidze

Elementary Probability Theory

2d edition, revised and completed

Moscow
2019

УДК 519.2(075.8)
ББК 22.171
Г18

Рецензенты: *Э.Л. Пресман*, д-р физ.-мат. наук, ЦЭМИ РАН
Н.О. Фастовец, канд. техн. наук, доцент, РГУ нефти и газа (НИУ) им. И.М. Губкина

ISBN 978–5-91961-393-0

Gamkrelidze N.G.

Elementary probability theory [Электронный ресурс] : учебное пособие. – 2d edition, revised and completed. – М. : РГУ нефти и газа (НИУ) имени И.М. Губкина, 2019. – 0,367 Мб. – Систем. требования: IBM-PC совместимый ; монитор ; привод CD-ROM ; программа для чтения PDF-файлов. – Загл. с тит. экрана.

Учебное пособие содержит элементарное введение в теорию вероятностей и рассчитано на широкий круг студентов, для которых вероятностные методы являются главным математическим инструментом. Пособие также окажется полезным всем тем, кто в дальнейшем предполагает изучать научную литературу на английском языке.

1 Preface

This text developed from an elementary course in probability theory at RSU. In preparing my lectures I borrowed heavily existing books and lectures in the field and the finished product reflects this. In particular the books H. Cramer, Yu. Prohorov and Yu. Rosanov, A. Shiryaev, P. Whittle were significant contributors.

2 Introduction

Probability is a mathematical science in which intuitive notions of "chance" or "randomness" are investigate. This one like all notions, is born of experience. Certain experiments are nonreproducible in that, when repeated under standard conditions, they produce variable results. The popular example is that of coin-tossing: the toss being the experiment, resulting in the observation of the number of heads $r(n)$. Following table shows a real result of this experiment.

Experiments by	Number of throws	Relative frequence of heads
Buffon	4040	0,5069
DeMorgan	4092	0,5005
K.Pearson	24000	0,5005

It is the empirical fact that $p(n) = \frac{r(n)}{n}$ varies with n much as in figure below.

The values of $p(n)$ show fluctuations which become progressively weaker as n increases, until ultimately $p(n)$ shows signs of tending to some kind of a limit value.

It is on this feature of empirical convergence that one founds probability theory; by postulating the existence of an idealized "proportion" (a probability) or "average" (an expectation).

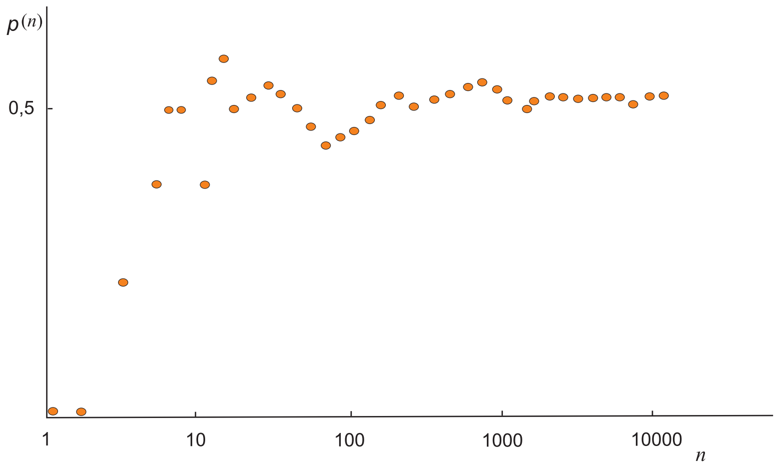


Fig. 1. A graph of the proportion of heads thrown, $p(n)$, in a sequence of n throws, from an actual coin-tossing experiment. Note the logarithmic scale for n.

It should be noted that $p(n)$ doesn't tends to its limit p in the usual sense of limits of sequence, because one cannot guarantee that the fluctuations in $p(n)$ will have fallen below a prescribed level for all values of n from a certain point onwards.

It is important to add that we proceed to work out a theory designed to serve as a mathematical model of phenomena showing statistical regularity.

Mathematical theory of probability don't investigate any uncertainty. Probability of truth that "there is a life on the another planet", or probability "following president of Russia will be a woman". Probabilities of this type have no direct connection with random experiments and so have not statistical regularity.

3 What is Elementary Probability Theory

A probabilistic model arising from the analysis of an experiment, the all possible result of which are expressed in a finite number of outcomes $\omega_1, ..., \omega_N$ is called an elementary probabilistic model and the corresponding theory is an elementary probability theory. We do not know the nature of these outcomes, only that there are finite number N of them.

Definition 1. The results of experiments or observations will be called *events*.

Definition 2. We call the finite set

$$\Omega = \{\omega_1, ..., \omega_N\}$$

the *space of elementary events or the sample space*.

Example 1. For a single toss of coin the space of elementary events Ω consists of two points:

$$\Omega = \{H, T\}$$

where $H =$ "head" $T =$ "tail".

Example 2. For n tosses of a coin the space of elementary events is

$$\Omega = \{\omega : \ \omega = \{\omega_1, ..., \omega_n\}\}, \qquad \omega_j = H \text{ or } T$$

and the general number $N(\Omega)$ of outcomes is 2^n.

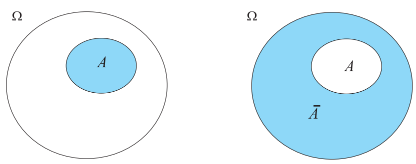


Fig. 2. A Venn diagram illustrating the complement $\bar{A}$ of a set A.
(Venn John (1834–1923) — English mathematicians).

Experimenters are ordinarily interested, not in what particular outcome occurs as the result of a trial but in whether the outcome belongs to some subset of the set of all possible outcomes.

Definition 3. We shall describe as *events* all subsets $A \subset \Omega$ for which, under the conditions of the experiment, it is possible to say either "the outcome $\omega \in A$" or the outcome $\omega \bar{\in} A$.

To every event there corresponds a opposite events "not A" to be denoted by $\bar{A}$. So if the event is "rain", then $\bar{A}$ is the event "no rain". *In set terms* $\bar{A}$ is the complement of A in Ω; the set of ω which does not lie in A.

Events are combined into new events by means of operations expressed by the terms "and" "or".

Definition 4. A "and" B is an event which occurs if, and only if, both the event A and the event B occur; denoted by $A \cap B$ or simply, AB. This is an intersection of A and B.

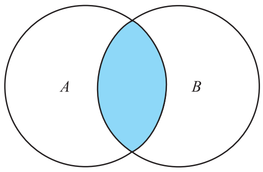


Fig. 3. A Venn diagram illustrating the intersection $A \cap B$ of sets A and B.

Suppose A and B are two events: Say "rain" and "wind" $A \cap B$ is the event that it rains and blows.

In set terms, the intersection $A \cap B$ is the set of ω belonging both to A and B.

Definition 5. A "or" B is an event which occurs if, and only if, at least on of the events A, B occurs, we denote it by $A \cup B$. This is a union of A and B. The union of events A — "rain" and B — "wind". That is, that it either rains or blows, or both.

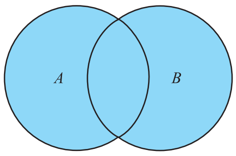


Fig. 4. A Venn diagram illustrating the union $A \cup B$ of sets A and B.

The set Ω is the set of all possible realizations, it is a *sure event*. Its complement is the empty set Θ, the set containing no elements at all, which can be referred to as a "never occurrence" and will be called the *impossible event*.

If events A and B are *mutually exclusive*, in that there is no realization for which they both occur, then the set $A \cap B$ is empty. That is, $A \cap B = \Theta$ and the set A and B are said to be disjoint. The difference $A \backslash B$ means that both A and $\bar{B}$ occur or, in other words, that A but not B occur: $A \cap \bar{B}$.

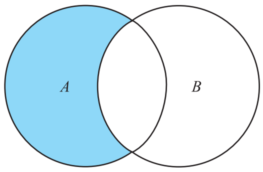


Fig. 5. A Venn diagram illustrating the difference $A \backslash B$ of sets A and $\bar{B}$.

3.1 Algebra of events

A collection $\mathcal{A}$ of subsets of Ω is an algebra if
(1) $\Omega \in \mathcal{A}$
(2) if $A \in \mathcal{A}$, $B \in \mathcal{A}$, than the sets $A \cup B$ (union), $A \cap B$ (intersection), $A \backslash B$ (difference) also belongs to $\mathcal{A}$.

Examples.

(a) $\{\Omega, \Theta\}$ trivial algebra

(b) $\{A, \bar{A}, \Omega, \Theta\}$, the collection generated by A

(c) $\mathcal{A} = \{A : \ A \subseteq \Omega\}$ the construction consisting of all the subsets of Ω (including the empty set Ω).

In elementary probability theory one usually takes the algebra $\mathcal{A}$ to be the algebra of all subsets of Ω.

3.2 Concept of probability

We have now taken the first two steps in defining a probabilistic model of an experiment with a finite number of outcomes: We have selected a sample space and a collection $\mathcal{A}$ of subsets, which form an algebra and are called events. We now take the next step, to assign to each sample point (outcome) $\omega_j \in \Omega \quad (j = 1, ..., N)$, a *weight*. This is denoted by $p(\omega_j)$ and called the probability of the outcome ω_j.

Starting from the given probabilities $p(\omega_j)$ of the outcomes ω_j, we define the probability $P(A)$ of any event $A \in \mathcal{A}$ by

$$P(A) = \sum_{\{j : \omega_j \in A\}} p(\omega_j).$$

In construction a probabilistic model for a specific situation, the constitution of the sample space Ω and the algebra $\mathcal{A}$ of events are ordinarily not difficult. Any difficulty that may arise is in assigning probabilities to the sample points. In the principle, the solution to this problem lies outside the domain of probability theory, and we shall not consider it in detail. We consider that our fundamental problem is not the question of how to assign probabilities, but how to calculate the probabilities of complicated events (element of $\mathcal{A}$) from the probabilities of the sample points. We assume that it has the following properties:

(1) Axiom of nonnegativity for any $A \in \mathcal{A} \qquad P(A) \geq 0$.

(2) Axiom of normalization $P(\Omega) = 1$.

(3) Axiom of additivity. If A and B are disjoint (mutually exclusive) sets (events): $A \cap B = \Theta$ then

$$P(A \cup B) = P(A) + P(B).$$

Finally, we say that a triple

$$(\Omega, \mathcal{A}, P)$$

where $\Omega = \{\omega_1, ..., \omega_n\}$, $\mathcal{A}$ is an algebra of subsets of Ω and

$$P = \{P(A); \qquad A \in \mathcal{A}\}$$

defines a probabilistic space.

In connection with the difficulty of assigning probabilities to outcomes, we note that there are many actual situations in which for reasons of symmetry it seems reasonable to consider all conceivable outcomes as equally probable. In such cases, if the sample space consists of points $\omega_1, ..., \omega_n$, with $n < \infty$, we put

$$p(\omega_1) = ... = p(\omega_n) = 1/n$$

and consequently

$$P(A) = n(A)/n,$$

for every event $A \in \mathcal{A}$, where $n(A)$ is the number of sample points in A. That is called the *classical method* of assigning probabilities. It is clear that in this case the calculation of $P(A)$ reduces to calculating the number of outcomes belonging to A. This is usually done by combinatorial methods, so that the combinatorics, applied to finite sets, plays a significant role in the calculus of probabilities. So we have following classical definition of mathematical probability. If there are n exhaustive, mutually exclusive and equally likely cases, and n_a of them are favorable to an event A the mathematical probability of A is defined as ratio n_A/n.

4 Conditional Probability. Independence

The concept of probabilities of events let us answer questions of the following kind: If there are M balls in an urn, M_1 white and M_2 black, what is the probability $P(A)$ of the event A that a selected ball is white? With the classical approach $P(A) = M_1/M$.

The concept of conditional probability, which will be introduced below, let us answer questions of the following kind: What is the probability that second ball is white (event B) under the condition that the first ball was also white (event A) (We think of sampling without replacement).

It is natural to reason as follows: If the first ball is white, then at the second step we have an urn containing $M - 1$ balls, of which $M_1 - 1$ are white and M_2 black; hence it seems reasonable to suppose that the (conditional) probability is $(M_1 - 1)/(M - 1)$.

We now give a definition of conditional probability that is consistent with our intuitive ideas.

Let $(\Omega, \mathcal{A}, P)$ be a finite probabilistic space and A an event (i. e. $A \in \mathcal{A}$).

Definition 1. The conditional probability of event B given event A with $P(A) > 0$ (denoted by $P_A(B)$) is

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

In the classical approach we have

$$P(A) = N(A)/N(\Omega), \quad P(A \cap B) = N(AB)/N(\Omega),$$

and therefore $P_A(B) = N(AB)/N(A)$.

From definition 1 we immediately get the following properties of conditional probability:

$$P_A(A) = 1, \;\; P_A(\Theta) = 0, \;\; P_A(B) = 1 \quad \text{if} \quad A \subset B,$$

$$P_A(B_1 + B_2) = P_A(B_1) + P_A(B_2).$$

Note that

$$P_A(B) + P_A(\bar{B}) = 1.$$

It follows from these properties that for a given event A the conditional probability $P_A(B)$ define probability distribution

values	$P(B/A_1)$	$P(B/A_2)$	...	$P(B/A_n)$
probability	$P(A_1)$	$P(A_2)$	...	$P(A_n)$

where $A_1 \cup ... \cup A_n = \Omega, \quad A_i \cap A_j = \Theta \quad \text{and} \quad P(A_k) > 0.$

Example. Consider a family with two children. We wan't to find the probability that both children are boys, assuming: a) that the older child is boy; b) that at least one of the children is a boy.

The sample space is $\Omega = \{BB, BG, GB, GG\}$, where BG means that older child is a boy and the younger is a girl.

Let us suppose that all sample points are equally probable

$$P(BB) = P(BG) = P(GB) = P(GG) = \frac{1}{4}.$$

Let A be the event that the older child is a boy, and C that the younger child is a boy. Then $A \cup C$ is the event that at least one child is a boy, and $A \cap C$ is the event that both children are boys. In question (a) we want to know the conditional probability $P_A(A \cap C)$, and in (b), the conditional probability $P_{A\cup C}(AC)$.

It is easy to see that

$$P_A(A \cap C) = \frac{P(A \cap C)}{P(A)} = \frac{1/4}{1/2} = \frac{1}{2}.$$

$$P_{A\cup C}(A \cap C) = \frac{P(A \cap C)}{P(A \cup C)} = \frac{1/4}{3/4} = \frac{1}{3},$$

because

$$A \cap C = BB, \qquad A \cup C = BB \cup BG \cup GB$$

and by axiom of additivity we have

$$\begin{aligned} P(A \cup C) &= P(BB \cup BG \cup GB) = \\ &= P(BB) + P(BG) + P(GB) = \tfrac{3}{4}. \end{aligned}$$

4.1 The formula for total probability

Definition. We say that the collection $D = \{D_1, ..., D_n\}$ of sets is a decomposition of Ω, and call the D_j the atoms of decomposition, if the D_j are not empty, are pairwise disjoint, and their sum is Ω.

Consider a decomposition

$$D = \{A_1, ..., A_n\} \quad \text{with} \quad P(A_j) > 0, \quad j = 1, ..., n$$

(such a decomposition is often called a complete set of disjoint events). It is clear that $B = BA_1 \cup ... \cup BA_n$ and since $BA_j \cap BA_k = \Theta$ $(j \neq k)$ disjoint events, we have

$$P(B) = P(BA_1 \cup ... \cup BA_n) = \sum_{j=1}^{n} P(BA_j).$$

But $P(BA_j) = P_{A_j}(B)P(A_j)$. Hence we have the formula for total probability

$$P(B) = \sum_{j=1}^{n} P_{A_j}(B)P(A_j).$$

In particular, if we take into account that $\Omega = A \cup \bar{A}$, then

$$P(B) = P_A(B)P(A) + P_{\bar{A}}(B) \cdot P(\bar{A}).$$

4.2 Bayes's formula

(Bayes Thomas (1702–1761) — English mathematician).

Suppose that A and B are events with $P(A) > 0$ and $P(B) > 0$. Then by the definition of conditional probability $P(A \cap B) = P_A(B)P(A)$. Then along with this formula we have the parallel formula $P(A \cap B) = P_B(A)P(B)$. From this formulae we obtain Bayes's formulaes

$$P_B(A) = \frac{P_A(B)P(A)}{P(B)}.$$

If the events $A_1, ..., A_n$ form a decomposition of Ω and we take into account the formula for total probability we have

$$P_B(A_j) = \frac{P_{A_j}(B)P(A_j)}{\sum\limits_{k=1}^{n} P(A_k)P_{A_k}(B)}$$ — this is *Bayes's theorem.*

Example. Let an urn contain two coins: A_1 is a fair coin with probability 1/2 of falling "H"; and A_2 — a biased coin with $P_2(H) = \frac{1}{3}$. A coin is drawn at random and tossed. Suppose that it falls head. We ask for the probability that the fair coin was selected.

Let us construct the corresponding space of elementary events $\Omega = \{A_1H, A_1T, A_2H, A_2T\}$, which describes all possible outcomes of a selection and a toss (A_1H means that coin A_1 was selected and fell heads).

$$P(A_1) = P(A_2) = 1/2,\ P_{A_1}(H) = 1/2,\ P_{A_2}(H) = 1/3.$$

Then by the definition 1 of conditional probability we have

$$P(A_1H) = P(A_1)P_{A_1}(H) = \tfrac{1}{2} \cdot \tfrac{1}{2} = \tfrac{1}{4}, \quad P(A_1T) = \tfrac{1}{4},$$
$$P(A_2H) = P(A_2)P_{A_2}(H) = \tfrac{1}{2} \cdot \tfrac{1}{3} = \tfrac{1}{6},$$
$$P(A_2T) = \tfrac{1}{2} \cdot \tfrac{2}{3} = \tfrac{1}{3}.$$

Using Bayes's formula, we get

$$P_H(A_1) = \frac{P(A_1)P_{A_1}(H)}{P(A_1)P_{A_1}(H) + P(A_2)P_{A_2}(H)} = \frac{3}{5}$$

and therefore $P_H(A_2) = \frac{2}{5}$.

4.3 Independence

Independence plays a central role in probability theory: it is precisely this concept that distinguishes probability theory from the general theory of measure spaces. "Probability theory is a measure theory — with a soul" (M. Kac).

After these preliminaries, we introduce the following definition.
Definition 1. Events A and B are called independent or statistically independent (with respect to the probability P) if

$$P(AB) = P(A)P(B)$$

Caution! Don't confuse notions of disjoint events and independence of events.

It is often convenient in probability theory to consider not only independence of events (or sets) but also independence of collections of events (or sets). Accordingly we introduce the following definition

Definition 2. Two algebras $\mathcal{A}_1$ and $\mathcal{A}_2$ of events (or sets) are called independent or statistically independent (with respect to the probability P) if all pairs of sets A_1, A_2 belonging respectively to $\mathcal{A}_1$ and $\mathcal{A}_2$, are independent.

Example 1. Let us consider two algebras

$$\mathcal{A}_1 = \{A_1, \bar{A}_1, \Theta, \Omega\} \quad \text{and} \quad \mathcal{A}_2 = \{A_2, \bar{A}_2, \Theta, \Omega\}$$

where A_1 and A_2 are subsets of Ω. It is easy to verify that $\mathcal{A}_1$ and $\mathcal{A}_2$ are independent if and only if A_1 and A_2 are independent. In fact, the independence of $\mathcal{A}_1$ and $\mathcal{A}_2$ means the independence of the 16 events A_1 and A_2, A_1 and $\bar{A}_2, ..., \Omega$ and Ω. Consequently A_1 and A_2 are independent. Conversely if A_1 and A_2 are independent, we have to show that the other 15 pairs of events are independent. Let us verify, for example, the independence of A_1 and $\bar{A}_2$. We have

$$P(A_1\bar{A}_2) = P(A_1(\Omega - A_2)) = P(A_1 - A_1A_2) =$$

$$= P(A_1) - P(A_1A_2) = P(A_1) - P(A_1)P(A_2) =$$

$$= P(A_1)(1 - P(A_2)) = P(A_1)P(\bar{A}_2).$$

The independence of the other pairs is verified similarly.

Example 2. A card is chosen at random from a deck of playing cards. For reasons of symmetry we expect the events "A=spade" and "B=ace" are independent. As a matter of fact, their probabilities are 1/4 and 1/13, and the probability of their simultaneous realization is

$$1/52 = P(AB) = P(A) \cdot P(B).$$

So we have statistically independent events!

Suppose now that three events A, B and C are pairwise independent so that

$$P(AB) = P(A) \cdot P(B), \quad P(AC) = P(A)P(C),$$
$$P(BC) = P(B)P(C).$$

We might think that this always implies the independence of A, B and C ($P(ABC) = P(A)P(B)P(C)$. Unfortunately this is not necessarily so!

Example 3 (Bernstein). Let us consider $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$ where all outcomes are equiprobable. It is easy to verify that the events $A = \{\omega_1, \omega_2\}, \quad B = \{\omega_1, \omega_3\}, \quad C = \{\omega_1, \omega_4\}$ are pairwise independent

$$P(AB) = P(A)P(B), \quad P(AC) = P(A)P(C),$$
$$P(BC) = P(B)P(C),$$

whereas

$$P(ABC) \neq P(A)P(B)P(C).$$

Indeed

$$P(ABC) = P(\omega_1) = \frac{1}{4} \quad \text{and} \quad P(A) = P(B) = P(C) = \frac{1}{2}.$$

It is desirable to reserve the term statistically independence for the case where no such inference is possible. Then not only pairwise independence must hold but in addition

$$P(ABC) = P(A) \cdot P(B) \cdot P(C).$$

Thus we have the following:

Definition 3. The events $A_1, A_2, ..., A_n$ are called mutually or statistically independent if for all combinations

$$1 \leq i < j < k < ... \leq n$$

the multiplication rules

$$P(A_iA_j) = P(A_i)P(A_j),$$

$$P(A_iA_jA_k) = P(A_i)P(A_j)P(A_k)$$

$$\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots$$

$$P(A_1A_2...A_n) = P(A_1) \cdot P(A_2)...P(A_n)$$

apply.

For the sake of simplicity we use in the following materials "independent" instead of "statistically independent". This concept can be extended to the any finite number of sets or algebras of sets: $\mathcal{A}_1, ..., \mathcal{A}_n$.

Definition 4. The algebras $\mathcal{A}_1, \mathcal{A}_2, ..., \mathcal{A}_n$ of events are called independent (with respect to the probability P) if all events $A_1, A_2, ..., A_n$ belonging respectively to $\mathcal{A}_1, ..., \mathcal{A}_n$ are independent.

Now let us consider following.

Example 4. Add to deck of playing cards a "white card". So the deck consists of 53 cards. We have

$$P(AB) = P(\text{"spade"} \cap \text{"ace"}) = \tfrac{1}{53}.$$

$$P(A = \text{"spade"}) = \tfrac{13}{53},$$

$$P(B = \text{"ace"}) = \tfrac{4}{53} \quad \text{and} \quad P(AB) \neq P(A)P(B).$$

So we conclude that the notions "Statistically independence" and "independence in every day sense" are different notions!

Now I would like to make an advance about following.

Caution! Don't confuse the notions "mutually independent" and "mutually exclusive" events!

Now let us answer following question, is there any object with condition of statistical independence?

4.4 Rademacher's function

(Rademacher Hans Adolf (1892–1969) — German mathematician).

The material of this section is useful but will not be used explicitly in the sequel.

It is well known that every number $x \in [0,1]$ has unique binary expansion (containing an infinite number of zeros)

$$x = \frac{\varepsilon_1}{2} + \frac{\varepsilon_2}{2^2} + \ldots \qquad (\varepsilon_i = 0,1).$$

For example

$$3/4 = 1/2 + 1/2^2 + 0/2^3 \ldots = (1,1).$$

To ensure the uniqueness of the expansion, we shall consider only the expansions containing an infinite number of zeros. Thus we choose the first of the two expansions

$$\frac{3}{4} = \frac{1}{2} + \frac{1}{2^2} + \frac{0}{2^3} + \ldots = \frac{1}{2} + \frac{0}{2^2} + \frac{1}{2^3} + \frac{1}{2^4} + \frac{1}{2^5} + \ldots$$

Let us take into account that ε_i is function of x i. e. $\varepsilon_i = \varepsilon_i(x)$ and consider $\varepsilon_1(x) = a_1 \quad (a_1 = 0,1)$. If $\varepsilon_1(x) = 0$, we have $x = \frac{\varepsilon_2(x)}{2^2} + \frac{\varepsilon_3(x)}{2^3} + \ldots$ and $x \in [0, \frac{1}{2})$. If $\varepsilon_1(x) = 1$ then $x = \frac{1}{2} + \frac{\varepsilon_2(x)}{2^2} + \ldots$ and $x \in [\frac{1}{2}, 1)$. See Fig. 6. So $P(x : \varepsilon_1(x) = a_1) = \frac{1}{2}$.

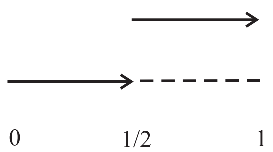


Fig. 6. The graph of the function $\varepsilon_1(x)$.

For $\varepsilon_2(x) = a_2 (a_2 = 0, 1)$

$$x = \frac{\varepsilon_1(x)}{2} + \frac{1}{2^2} + \frac{\varepsilon_3(x)}{2^3} + ... \quad \text{if} \quad \varepsilon_2(x) = 1$$

or

$$x = \frac{\varepsilon_1(x)}{2} + \frac{\varepsilon_3(x)}{2^3} + ... \quad \text{if} \quad \varepsilon_2(x) = 0.$$

So, if $\varepsilon_2(x) = 1$ then $x \in [\frac{1}{4}, \frac{1}{2})$ or $x \in [\frac{3}{4}, 1)$ and $P(x : \varepsilon_2(x) = 1) = \frac{1}{2}$. Similarly if $\varepsilon_2(x) = 0$ then $x \in [0, \frac{1}{4})$ or $x \in [\frac{1}{2}, \frac{3}{4})$ (See Fig. 7). So $P(x : \varepsilon_2(x) = 0) = \frac{1}{2}$.

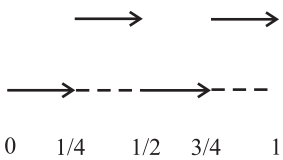


Fig. 7. The graph of the function $\varepsilon_2(x)$.

Now let us consider the following expression

$$P(x : \varepsilon_1(x) = a_1, \varepsilon_2(x) = a_2) =$$

$$= P(x : \tfrac{a_1}{2} + \tfrac{a_2}{2^2} \leq x < \tfrac{a_1}{2} + \tfrac{a_2}{2} + \tfrac{1}{2^2}).$$

Assume that $a_1 = 0, \ a_2 = 0$, then

$$\begin{aligned}
P(x : \varepsilon_1(x) &= 0, \ \varepsilon_2(x) = 0) = P(x : 0 \leq x \leq \tfrac{1}{4}) = \tfrac{1}{4}; \\
P(x : \varepsilon_1(x) &= 0, \ \varepsilon_2(x) = 1) = P(x : \tfrac{1}{2^2} \leq x < \tfrac{1}{2^2} + \tfrac{1}{4}) = \tfrac{1}{4}; \\
P(x : \varepsilon_1(x) &= 1, \ \varepsilon_2(x) = 0) = P(x : \tfrac{1}{2} \leq x < \tfrac{1}{2} + \tfrac{1}{4}) = \tfrac{1}{4}; \\
P(x : \varepsilon_1 &= 0, \ \varepsilon_2(x) = 1) = P(x : \tfrac{1}{2^2} \leq x < \tfrac{1}{2^2} + \tfrac{1}{4}) = \tfrac{1}{4}.
\end{aligned}$$

This means that $P(x:\ \varepsilon_1(x)=a_1,\quad \varepsilon_2(x)=a_2)=\frac{1}{4}$. Therefore

$$\frac{1}{4}=t(x:\varepsilon_1(x)=a_1,\quad \varepsilon_2(x)=a_2)=$$
$$=P(x:\ \varepsilon_1(x)=a_1)\cdot P(x:\ \varepsilon_2(x)=a_2)=\frac{1}{4}.$$

This establishes that $\varepsilon_1(x)$ and $\varepsilon_2(x)$ are statistically independent. The same is true for $\varepsilon_1(x),...,\varepsilon_n(x)$.

If we now set $R_n(x)=1-2\varepsilon_n(x)$, $n\geq 1$ that the sequence, so called Rademacher's function $R_1(x)$, $R_2(x)$,..., $R_n(x)$ are independent!

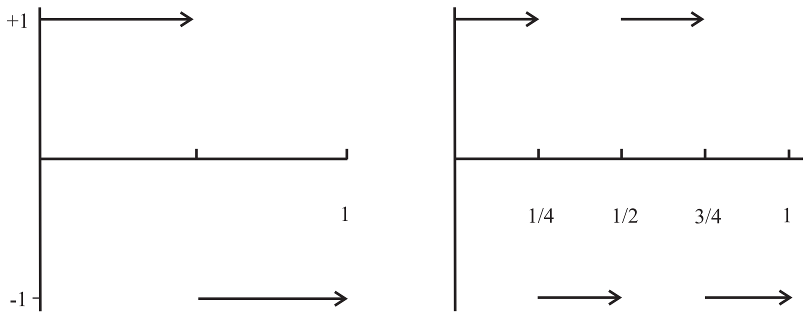


Fig. 8. The graphs of the functions $R_1(x)$ and $R_2(x)$.

5 Random Variables and their Properties

The concept "random variable", which we now introduce serves to define the quantities that are subject to "measurement" in random experiments.

Definition. Any numerical function $\xi=\xi(\omega)$ defined on a (finite) sample space Ω is called a (simple) random variable. (The reason for the term "simple" random variable will become clear after the introduction of the general concept of random variable).

Example. In the model of two tosses of a coin with sample space $\Omega = \{HH, HT, TH, TT\}$, define a random variable $\xi = \xi(\omega)$ by the table

ω	HH	HT	TH	TT
$\xi(\omega)$	2	1	1	0

Here, from its very definition, $\xi(\omega)$ is nothing but the number of heads in the outcome ω.

Another simple example of a random variable is the indicator function of a set $A \in \mathcal{A} : \xi = I(A)$ where

$$I(A) = I_A(\omega) = \begin{cases} 1, & \text{if} \quad \omega \in A \\ 0, & \text{if} \quad \omega \bar{\in} A. \end{cases}$$

When experiments are concerned with random variables that describe observations, their main interest is in the probabilities with which the random variables take various values. Since we are considering the case when Ω contains only a finite number of points, the range x of the random variable ξ is also finite. Let $X = \{x_1, ..., x_m\}$ where the (different) numbers $x_1, ..., x_m$ exhaust the values of ξ. If we put

$$A_j = \{\omega : \ \xi(\omega) = x_j\} \qquad (j = 1, 2, ..., m),$$

then ξ can evidently be represented by as

$$\xi = \xi(\omega) = \sum_{j=1}^{m} x_j I_{A_j}(\omega),$$

where the sets $A_1, ..., A_m$ form a decomposition of Ω; (i.e. they are pairwise disjoint and their sum is Ω. See 4.1). It is clear that the values of $P_\xi(B) = P\{\omega : \xi(\omega) \in B\}$, $B \in \mathfrak{X}$ ($\mathfrak{X}$ be the collection of all subsets of X) are completely determined by the probabilities $P_\xi(x_j) = P\{\omega : \xi(\omega) = x_j\}$ $x_j \in X$.

Definition. The set of numbers

$$\{P_\xi(x_1), ..., P_\xi(x_m)\}$$

is called the probability distribution of the random variable ξ.

Example. A random variable ξ that takes the two volumes 1 and 0 with probabilities p ("success") and q ("failure") is called a Bernoulli random variable. Clearly $P_\xi(x) = p^x q^{1-x}$, $x = 0, 1$. A binomial (or binomially distributed) random variable ξ is a random variable that takes the $n+1$ values $0, 1, ..., n$ with probabilities $P_\xi(x) = C_n^x p^x q^{n-x}$, $x = 0, 1, ..., n$. The probabilistic structure of the random variables ξ is completely specified by the probability distribution

$$\{P_\xi(x_j), j = 1, ..., m\}.$$

The concept of the distribution function, which we now introduce, yields an equivalent description of the probabilistic structure of the random variables.

Definition. Let $x \in R^1$. The function

$$F_\xi(x) = P\{\omega\colon\ \xi(\omega) \leq x\}$$

is called the distribution function of the random variable ξ.

Clearly

$$F_\xi(x) = \sum_{j:x_j \leq x} P_\xi(x_j)$$

and $P_\xi(x_j) = F_\xi(x_j) - F\xi(x_j-)$ where

$$F_\xi(x-) = \lim_{y \uparrow x} F_\xi(y).$$

Here $\lim_{y\uparrow x} F_\xi(y)$ is the left-hand limit. If we suppose that $x_1 < x_2 < ... < x_m$ and put $F_\xi(x_0) = 0$, then

$$P_\xi(x_j) = F_\xi(x_j) - F_\xi(x_{j-1}), \ \ (j = 1, ..., m).$$

$P_j = P_\xi(x_j)$ — probability distribution, $F_\xi(x)$ — distribution function.

It follows from the last Definition that the distribution function $F_\xi(x)$ has the following properties:

1. $F_\xi(-\infty) = 0; \qquad F_\xi(\infty) = 1.$

2. $F_\xi(x)$ is continuous on the right $F_\xi(x+0) = F_\xi(x)$, and $F_\xi(x)$ has left-hand limit. The function $F_\xi(x)$ is "CADLAG"[1].

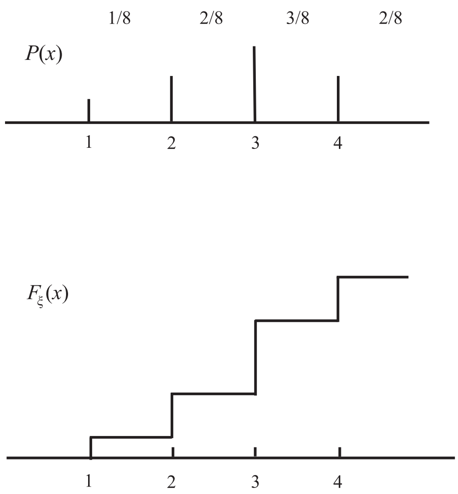


Fig. 9. The graphs of $P(x)$ and $F_\xi(x)$ for a random variable ξ with values $1, 2, 3, 4$ and probabilities $\frac{1}{8}, \frac{2}{8}, \frac{3}{8}, \frac{2}{8}$ accordingly.

We now turn to the important concept of independence of random variables.

Let $\xi_1, ..., \xi_r$ be a set of random variables with values in a finite set $X \subseteq R^1$.

Definition. The random variables $\xi_1, ..., \xi_r$ are said to be *independent* (mutually independent) if

$$P(\xi_1 = x_1, ..., \xi_r = x_r) = P(\xi_1 = x_1) \cdot ... \cdot P(\xi_r = x_r)$$

for all $x_1 ... x_r \in X$.

[1]The term "CADLAG" is an acronym for the French phrase which means "continuous on the right, limits on the left".

6 The Binomial Distribution

Let a coin be tossed n times and record the results as an ordered set $(a_1, ..., a_n)$, where $a_j = 1$ for a head ("success") and $a_j = 0$ for a tail ("failure"). The space of elementary events is

$$\Omega = \{\omega : \ \omega = (a_1, ..., a_n), \ a_j = 0, 1).$$

To each elementary event $\omega = (a_1, ..., a_n)$ we assign the probability

$$p(\omega) = p^{\sum a_j} q^{n - \sum a_j},$$

where nonnegative numbers p and q satisfy $p + q = 1$. We consider all outcomes $\omega = (a_1, ..., a_n)$ for which

$$\sum_j a_j = k \qquad (k = 0, 1, ..., n).$$

According to that (the distribution of k indistinguishable ones in places) the number of these outcomes is C_n^k. Therefore the binomial formula gives

$$\sum_{\omega \in \Omega} p(\omega) = \sum_{k=0}^{n} C_n^k p^k q^{n-k} = (p+q)^n = 1.$$

So we verify that this assignment of the weights $p(\omega)$ is consistent because we show that

$$\sum_{\omega \in \Omega} p(\omega) = 1.$$

Thus the space Ω together with the collection $\mathcal{A}$ of all its subsets and the probabilities

$$P(A) = \sum_{\omega \in A} p(\omega), \qquad A \in \mathcal{A},$$

defines a probabilistic model for n tosses of a coin. We note that this model for n tosses of a coin can be thought of as the result of

n "independent" experiments with probability p of success at each trial.

Let us consider the events

$$A_k = \{\omega : \omega = (a_1, ..., a_n), \;\; a_1 + ... + a_n = k\}, \;\; k = 0, 1, ..., n,$$

consisting of exactly k successes. It follows from what we said above that $P(A_k) = C_n^k p^k q^{n-k}$, and

$$\sum_{k=0}^{n} p(A_k) = 1.$$

Definition. The set of probabilities $(P(A_0), ..., P(A_n))$ is called the binomial distribution (the number of successes in a n tosses).

This distribution plays an extremely important role in probability theory since it arises in the most diverse probabilistic models. We write $P_n(k) = P(A_k), \;\; k = 0,1, ...n$. The following figure shows the binomial distribution in the case $p = \frac{1}{2}$ (symmetric coin) for $n = 10$.

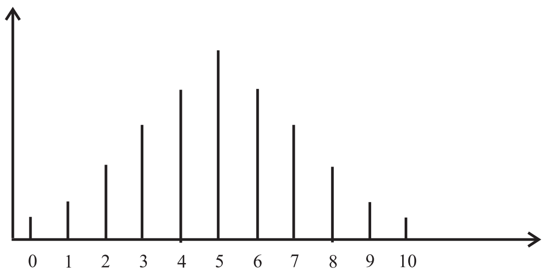


Fig. 10. The graph of the binomial distribution in the case $p = \frac{1}{2}$ for $n = 10$.

6.1 The Poisson distribution

Definition. The set of probabilities $P(A_1)P(A_2), ..., P(A_n)...$ where $P(A_k) = \frac{\lambda^k e^{-\lambda}}{k!}$ is called the Poisson distribution.

This is indeed probability distribution because

$$\sum_{k=1}^{\infty} \frac{\lambda^k e - \lambda}{k!} = e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^k}{k!} = e^{-\lambda} \cdot e^{\lambda} = 1!$$

Notice that all the (discrete) distributions considered previously were concentrated at only a finite number of points. The Poisson distribution is the first example that we have encountered of discrete distribution concentrated at a countable number of points.

6.2 Poisson's theorem

(Poisson Simeon Deni (1781–1840) — French mathematician).

Let

$$P_n(k) = \begin{cases} C_n^k p^k q^{n-k}, & k = 0, 1, 2, ..., n \\ 0 & k = n+1, n+2, ... \end{cases}$$

and suppose p is a function $p(n)$ of n.

Theorem. Let $p(n) \to 0$, $n \to \infty$ in such a way that $np(n) \to \lambda$, where $\lambda > 0$, $(\lambda < \infty)$. Then for $k = 1, 2, ...$

$$P_n(k) \to \pi_k = \frac{\lambda^k e^{-\lambda}}{k!}$$

$$|P_n(k) - \pi_k| \leq \frac{3}{2}\frac{\lambda^2}{n}.$$

Proof.

$$P_n(k) = C_n^k p^k q^{n-k} =$$

$$= \tfrac{n!}{k!(n-k)!}[\tfrac{\lambda}{n} + \mathrm{o}(\tfrac{1}{n})]^k \cdot [1 - \tfrac{\lambda}{n} + \mathrm{o}(\tfrac{1}{n})]^{n-k} =$$

$$= \tfrac{n(n-1)\cdot\ldots\cdot(n-k+1)}{k!}[\tfrac{\lambda}{n} + \mathrm{o}(\tfrac{1}{n})]^k[1 - \tfrac{\lambda}{n} + \mathrm{o}(\tfrac{1}{n})]^{n-k}$$

But

$$n(n-1)\cdot ... \cdot(n-k+1)[\tfrac{\lambda}{n}+0(\tfrac{1}{n})]^k =$$

$$= \tfrac{n(n-1)\cdot ... \cdot(n-k+1)}{n^k}[\lambda+0(1)]^k =$$

$$= (1-\tfrac{1}{n})\cdot ... \cdot(1-\tfrac{k+1}{n})[\lambda+0(1)]^k \to \lambda^k$$

as $n\to\infty$, and

$$[1-\frac{\lambda}{n}+0(\frac{1}{n})]^{n-k}\to e^{-\lambda}$$

as $n\to\infty$, which establishes this theorem.

In the preceding theorem we have used the Poisson distribution merely as a convenient approximation to the binomial distribution in the case of large n and small p. In many applications we deal with Poisson distribution as a principal distribution of probability theory. Stars in space, raisins in cake, misprints are distributed in accordance with the Poisson law!

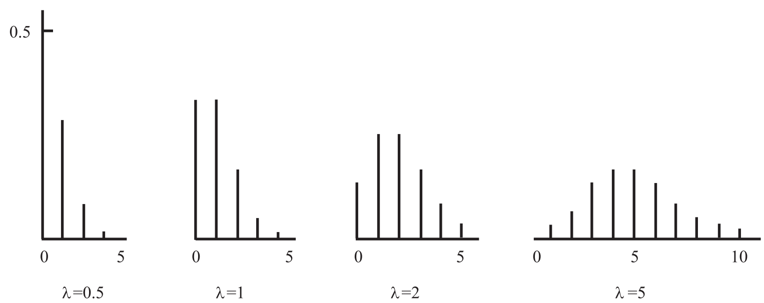


Fig. 11. The Poisson distribution for various values of λ.

7 The Hypergeometric Distribution

Consider, for example, an urn containing M balls numbered $1, 2, ..., M$ where M_1 balls have the color b_1 and M_2 balls have the color b_2 and $M_1 + M_2 = M$. Suppose that we draw a sample of size $n < M$ without replacement. The sample space is $\Omega =$

$\{\omega : \ \omega = (a_1, ..., a_n)\} \ a_k \neq a_j, \ k \neq j, \ a_j = 1, ..., M\}$ and the number of elementary events $N(\omega)$ is equal C_M^n. Let us suppose that the space of elementary events are equiprobable and find the probability of the event $B_{n_1 n_2}$ in which n_1 balls have color b_1 and n_2 balls have color b_2, where $n_1 + n_2 = n$. It is easy to show that $N(B_{n_1 n_2}) = C_{M_1}^{n_1} C_{M_2}^{n_2}$.

Definition. The set of probabilities $\{P(B_{n_1,n_2} \ M_1 + M_2 = M)\}$ is called the hypergeometric distribution:

$$P(B_{n_1,n_2}) = \frac{C_{M_1}^{n_1} C_{M_2}^{n_2}}{C_M^n}; \quad (n_1 + n_2 = n).$$

Example. 1) Let us consider a lottery of the following kind. There are 50 balls numbered from 1 to 50; 7 of them are lucky. We draw a sample of 7 balls, without replacement. The person who picks 7 "lucky" numbers wins a billion$^{-}10^9$ dollars! The question is what is the probability of this events? Taking $M = 50, \ \ M_1 = 7, \ \ n_1 = 7, \ \ n_2 = 0$.

$$\begin{aligned} P(B_{7,0}) &= P(7 \text{ balls, all lucky}) = \\ &= \tfrac{C_7^7 \cdot C_{43}^0}{C_{50}^7} = \tfrac{1}{C_{50}^7} \simeq 2{,}33 \cdot 10^{-10}. \end{aligned}$$

2) Sportloto. There are 49 balls numbered from 1 to 49. 6 of them are lucky. What is the probability of the events: $P(B_{j,k}) = P$ (between 6 balls: j lucky, k unlucky) where $j, k = 0, 1, ..., 6$ and $j + k = 6$.

Taking $M = 49, \ M_1 = 6$:

1) $n_1 = 6 \ \ n_2 = 0 \ \ P(B_{6,0}) = \frac{C_6^6 C_{49^6}^0}{C_{49}} = 7{,}2 \cdot 10^{-8}$;

2) $P(B_{5,1}) = Pr(\text{between 6 balls: 5 lucky, 1 unlucky}) = $
$= \frac{C_6^5 C_{43}^1}{C_{49}^6} = 0{,}00001858$;

3) $P(B_{4,2}) = Pr(\text{between 6 balls: 4 lucky, 2 unlucky}) = $
$= \frac{C_6^4 C_{43}^2}{C_{49}^6} = 0{,}000969$;

4) $P(B_{3,3}) = Pr(\text{between 6 balls: 3 lucky, 3 unlucky}) = $
$= \frac{C_6^3 C_{43}^3}{C_{46}^6} = 0{,}017650$;

5) $P(B_{2,4}) = Pr(\text{between 6 balls: 2 lucky, 4 unlucky}) =$
$$= \frac{C_6^2 C_{43}^4}{C_{49}^6} = 0{,}132378;$$

6) $P(B_{1,5}) = Pr(\text{between 6 balls: 1 lucky, 5 unlucky}) =$
$$= \frac{C_6^1 C_{43}^5}{C_{49}^6} = 0{,}413019;$$

7) $P(B_{0,6}) = Pr(\text{between 6 balls: 0 lucky, 6 unlucky}) =$
$$= \frac{C_6^0 C_{43}^6}{C_{49}^6} = 0{,}435965.$$

To make this example clear let us consider an urn with 5 balls numbered $1, 2, ..., 5$. 3 of them are white, 2 — black. What is the probability Pr (between 3 balls: 2 white and 1 black)=? $Pr(\cdot) = \frac{C_3^2 C_2^1}{C_5^3}$;

8 The Continuous Type of Distribution

Before this paragraph we had to deal with discrete probabilities and it is possible approximations of the following form

$$P(a < \xi < b) \approx \int\limits_a^b f(x)dx.$$

In many cases this passage to the limit leads conceptually to a new — continuous space of elementary events, and the latter may be intuitively simples than the original discrete model but the definition of probabilities in it depends on tools such as integration and measure theory.

Example (Feller). Random choices. To "choose a point at random" in the interval (0,1) is a conceptual experiment with an obvious intuitive meaning. It can be described by discrete approximations, but it is easier to use the whole interval as an sample space of events and to assign to each interval its length as probability. The conceptual experiment of making two independent random variable choice of points in (0,1) results in a pair of real numbers, and so the natural space of elementary events is a unit square. In this sample space of elementary events one equates, almost instinctively "probability" with "area". This is quite satisfactory for some elementary purpose, but sooner or later the question arises as to what the word "area" really means.

Definition. A variable ξ will be said to be of the continuous type, or to possess a distribution of this type, if the distribution function $F(x) = P(\xi \leq x)$ is everywhere continuous and if, the derivative $F'(x) = f(x)$ exists in a certain point x, then we shall call $f(x)$ the probability density function. Moreover, if the density function $f(x) = F'(x)$ is continuous for all values of x_1 except possibly in certain points of which any finite interval contains at most a finite number. The distribution function $F(x)$ is then

$$F(x) = P(\xi \leq x) = \int_{-\infty}^{x} f(t)dt.$$

The distribution has no discrete mass points, and consequently the probability that ξ takes a value x_0 is zero for every x_0: $P(\xi = x_0) = 0$. The probability that ξ takes a value belonging to the finite or infinite interval (a, b) has thus the same value, whether we consider the interval as closed, open or half-open and is given by

$$P(a < \xi < b) = F(b) - F(a) = \int_{a}^{b} f(t)dt.$$

Since the total mass in the distribution must be unity, we always have

$$\int_{-\infty}^{\infty} f(t)dt = 1.$$

A distribution of the continuous type may be graphically represented by diagrams, showing distribution function $F(x)$ or the density function $f(x)$.

9 The Normal Distribution

Definition. The function defined by

$$\varphi(x, m, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-m)^2/2\sigma^2}$$

is called the normal density function, its integral

$$\Phi\left(\frac{x-m}{\sigma}\right) = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{\frac{x-m}{\sigma}} e^{-\frac{u^2}{2}} du = \frac{1}{\sigma\sqrt{2\pi}} \int\limits_{-\infty}^{x} e^{-\frac{(u-m)^2}{2\sigma^2}} du,$$

is the normal distribution function with parameters m and σ. The transformation $u = \frac{x-m}{\sigma}$ carries the normal law with parameters m and σ into the *standard normal distribution* with parameters $m = 0$, $\sigma = 1$ and density

$$\frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} \qquad (-\infty < u < \infty).$$

This distribution plays an exceptionally important role. This comes about, first of all, because under rather general hypotheses, sums of a large number of independent random variables are closely approximated by normal distribution.

The $\varphi(x)$ is a symmetric bell-shaped curve, decreasing very rapidly with increasing $|x|$:

$$\begin{aligned} \varphi(1) &= 0{,}24197, \\ \varphi(2) &= 0{,}053991, \\ \varphi(3) &= 0{,}004432, \\ \varphi(4) &= 0{,}000134, \end{aligned}$$

the graph of which is shown in Fig. 12.

The curve $\Phi(x)$ approximates 1 very rapidly as x increases:

$$\begin{aligned} \Phi(1) &= 0{,}841345, \\ \Phi(2) &= 0{,}977250, \\ \Phi(3) &= 0{,}998650, \\ \Phi(4) &= 0{,}999968. \end{aligned}$$

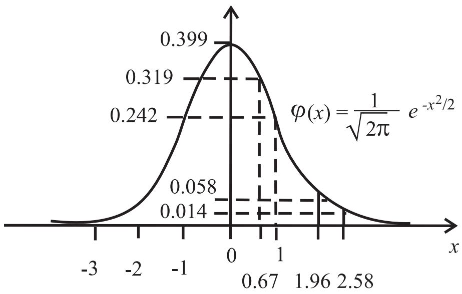


Fig. 12. The graph of the normal probability density $\varphi(x)$.

For tables $\varphi(x)$ and $\Phi(x)$, as well as of other functions that are used in probability theory. (See: [1], [7]).

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^2 e^{-t^2}\, dt;\ \Phi(x) = \frac{1}{2}[1 + \operatorname{erf}(\frac{x}{\sqrt{2}})].$$

10 Expectation

We now turn to the random variable with finite number of values.

Let $p_i = P\{\xi = x_i\}$. It is intuitively plausible that if we observe the values of the random variable ξ in "n repetitions of identical experiments", the value x_j ought to be encountered about $p_j n$ times, $j = 1, ..., k$. Hence the mean value calculated from the results of n experiments is roughly

$$\frac{1}{n}[np_1x_1 + ... + np_kx_k] = \sum_{j=1}^{n} p_j x_j.$$

This discussion provides the motivation for the following definition.

Definition. The expectation (or mathematical expectation) or mean value of the random variable

$$\xi = \sum_{j=1}^{n} x_j I(A_j)$$

is the number

$$E\xi = \sum_{j=1}^{n} x_j P(A_j)$$

where

$$A_j = \{\omega : \xi(\omega) = x_j\}, \quad \bigcup_{j=1}^{n} A_j = \Omega \quad \text{and} \quad A_i \cap A_j = \Theta.$$

Since $P_\xi(x_j) = P(A_j)$, we have

$$\boxed{E\xi \overset{def}{=} \sum_{j=1}^{n} x_j P_\xi(x_j).}$$

Reminder.

$$I(A_j) = \begin{cases} 1 & \omega_j \in A_j = \{\omega : \ \xi(\omega) = x_j\} \\ 0 & \omega_j \in \bar{A}_j = \{\omega : \ \xi(\omega) \neq x_j\} \end{cases}$$

$$EI(A_j) = 1 \cdot P(A_j) + 0 \cdot P(\bar{A}_j) = P(A_j).$$

We list the basic properties of expectation:

1. If $\xi \geq 0$, then $E\xi \geq 0$. This property is evident.
2. If ξ and η are arbitrary random variables, then,

$$E(a\xi + b\eta) = aE\xi + bE\eta,$$

where a and b are constants.

Let

$$\xi = \sum_j x_j I(A_j), \qquad \eta = \sum_j y_j I(B_j);$$

Then

$$\begin{aligned} a\xi + b\eta &= a\sum_{i,j} x_j I(A_j \cap B_i) + b\sum_{i,j} y_j I(A_i \cap B_j) = \\ &= \sum_{i,j}(ax_j + by_i) I(A_j \cap B_i) \end{aligned}$$

and

$$\begin{aligned} E(a\xi + b\eta) &= \sum_{i,j}(ax_i + by_i) P(A_i \cap B_j) = \\ &= \sum_i ax_i P(A_i) + \sum_j by_j P(B_j) = aE\xi + bE\eta. \end{aligned}$$

So we have

$$\boxed{E(a\xi + b\eta) = aE\xi + bE\eta}$$

Particularly $Ea = a$.

3. If $\xi \geq \eta$ then $E\xi \geq E\eta$. This property follows from 1. and 2.

4. $\mid E\xi \mid \leq E \mid \xi \mid$. This is evident, since

$$\mid E\xi \mid = \mid \sum_j x_j P(A_j) \mid \leq \sum_j \mid x_j \mid P(A_j) = E \mid \xi \mid .$$

5. If ξ and η are independent, then $E(\xi\eta) = E\xi \cdot E\eta$.

To prove this we note that

$$\begin{aligned} E(\xi\eta) &= E[(\sum_j x_j I(A_j)) \cdot (\sum y_i I(B_j)] = \\ &= E(\sum_{i,j} x_i y_j I(A_i \cap B_j)) = \sum_{i,j} x_i y_i P(A_i \cap B_j) = \\ &= \sum_{i,j} x_i y_j P(A_i) P(B_j) = \\ &= (\sum_i x_i P(Ai))(\sum_j y_j P(B_j)) = E\xi \cdot E\eta. \end{aligned}$$

Remark. If $E\xi\eta = E\xi E\eta$ it does not follow in general that they are independent: $P(\xi = x)P(\eta = y) = P(\xi = x, \eta = y)$. Let us consider random variable α which takes the values $0, \pi/2, \pi$ with probability $1/3$:

α	0	$\pi/2$	π
	$1/3$	$1/3$	$1/3.$

Then $\xi = \sin\alpha$ and $\eta = \cos\alpha$.

$$E\xi = E\sin\alpha = \frac{1}{3}(\sin 0 + \sin\frac{\pi}{2} + \sin\pi) = \frac{1}{3}.$$

$$E\eta = E\cos\alpha = \frac{1}{3}(\cos 0 + \cos\frac{\pi}{2} + \cos\pi) = 0.$$

$$\begin{aligned} E(\eta\xi) &= E(\sin\alpha\cos\alpha) = \tfrac{1}{2}E\sin 2\alpha = \\ &= \tfrac{1}{2}\cdot\tfrac{1}{3}(\sin 0 + \sin\pi + \sin 2\pi) = 0 \end{aligned}$$

So, $E\xi \cdot E\eta = \frac{1}{3}\cdot 0 = 0 = E(\eta\xi)$.

$$P(\xi = 1) = \tfrac{1}{3}\cdot\sin\tfrac{\pi}{2} = \tfrac{1}{3},$$
$$P(\eta = 1) = \tfrac{1}{3}\cos 0 = \tfrac{1}{3} \qquad P(\xi = 1 \cap \eta = 1) = 0$$

and

$$0 = P(\xi = 1 \cap \eta = 1) = P(\xi = 1)\cdot P(\eta = 1) = \frac{1}{9}!$$

So,

$$E\xi\eta = E\xi E\eta \not\Longrightarrow P(\xi = x)P(\eta = y) = P(\xi = x, \eta = y).$$

6. $(E \mid \xi\eta \mid)^2 \leq E\xi^2 \cdot E\eta^2$ (Cauchy–Bunyakovsky–Schwarz inequality without proof! See [14]).

7. If $\xi = I(A)$ then $E\xi = P(A)$, by definition $I(A)$ we have

$$I(A) = \begin{cases} 1 & \omega \in A \\ 0 & \omega \in \bar{A} \end{cases}$$

$$E\xi = EI(A) = 1\cdot P(A) + 0\cdot P(\bar{A}) = P(A),$$

where

$$A = \{\omega : \xi(\omega) = 1\}, \qquad \bar{A} = \{\omega : \xi(\omega) = 0\}.$$

8. Let $\xi = \sum\limits_j x_j I(A_j)$, where $A_j = \{\omega : \ \xi(\omega) = x_j\}$, and $\varphi = \varphi(\xi(\omega))$ is a function of $\xi(\omega)$. If $B_j = \{\omega_j \ \varphi(\xi(\omega)) = y_j\}$, then

$$\varphi(\xi(\omega)) = \sum_j y_j I_\varphi(B_j), \quad \text{then}$$

$$I(B_j) = I_\varphi(B_j) = \begin{cases} 1 & \varphi = y_j \\ 0 & \varphi \neq y; \end{cases}$$

and consequently

$$E\varphi(\xi) = \sum_j y_j P(B_j) = \sum_j y_j P(\varphi(y_j)),$$

where

$$P(\varphi(y_j)) = P\{\omega : \varphi(\xi(\omega) = x_j) = y_j\}.$$

Hence the expectation of the random variable $\varphi = \varphi(\xi)$ can be calculated as

$$E\varphi(\xi) = \sum_j \varphi(x_j) P_\xi(x_j),$$

where

$$P_\xi(x_j) = P\{\omega : \xi(\omega) = x_j\}.$$

Exercise. Let the random variable ξ take the values $0, 10$ with probability $1/2$

$$\xi : \begin{array}{c|c|c} x_j & 0 & 10 \\ \hline P_j & 1/2 & 1/2 \end{array}$$

Find the expectation of ξ.

$$E\xi = 0 \cdot \frac{1}{2} + 10\frac{1}{2} = \frac{10}{2} = 5$$

Example. Let ξ be a Bernoulli random variable, taking the values 1 and 0 with probabilities p and q ($p + q = 1$). Then $E\xi = 1 \cdot P(\xi = 1) + 0 \cdot P(\xi = 0) = p$.

Example. Let $\xi_1, ..., \xi_n$ be n Bernoulli random variables with $P(\xi_j = 1) = p$, $P(\xi_j = 0) = q$, $p + q = 1$. Then if $S_n = \xi_1 + ... + \xi_n$ we find $ES_n = E\xi_1 + ... + E\xi_n = np$.

Example. Let ξ be a Poisson random variable

$$P(\xi = m) = \frac{\lambda^m}{m!} e^{-\lambda}.$$

Then

$$E\xi = \sum_{m=1}^{\infty} m\frac{\lambda^m}{m!}e^{-\lambda} = \lambda e^{-\lambda}\sum_{m=1}^{\infty}\frac{\lambda^{m-1}}{(m-1)!} = \lambda e^{-\lambda}e^{\lambda} = \lambda.$$

Because

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

10.1 Conditional expectations

Given a probability space (Ω, S, P) and two events A and B in S with $P(B) > 0$

Definition. The conditional probability of A given B is defined as

$$P(A/B) := P(A \cap B)/P(B).$$

The conditional expectation of random variable X given the event B is defined (when it exists) as

$$E(X/B) := \left(\int_B X dP\right)/P(B).$$

Martingales. Suppose $\{B_t\}_{t\in T}$ is a family of σ-algebras with $B_t \subset B_n \subset B$ for $t \leq n$. Then $\{X_t, B_t\}$ is called a *martingale* iff $E|X_t| < \infty$ for all t and $X_t = E(X_n/B_t)$ whenever $t \leq n$

If we think of X_t as the fortune at time t of a gambler, then a martingale is "fair" game in the sense that at any time t, no matter the history up to the present (given by B_t), the expected net gain or loss from further play to time t is 0.

11 Variance

Definition. The *variance* (or dispersion) of the random variable ξ (denoted by $V\xi$) is

$$\boxed{V\xi \stackrel{def}{=} E(\xi - E\xi)^2.}$$

The number $\sigma \stackrel{def}{=} +\sqrt{V\xi}$ is called the *standard deviation*. Since

$$\begin{aligned} E(\xi - E\xi)^2 &= E\left(\xi^2 - 2\xi E\xi + (E\xi)^2\right) = \\ &= E\xi^2 - 2E\xi \cdot E\xi + E(E\xi)^2 = \\ &= E\xi^2 - 2(E\xi)^2 + (E\xi)^2 = E\xi^2 - (E\xi)^2, \end{aligned}$$

we have

$$V\xi = E\xi^2 - (E\xi)^2.$$

Clearly $V\xi \geq 0$. It follows from definition that

$$\begin{aligned} V(a + b\xi) &= E[a + b\xi - E(a + b\xi)]^2 = \\ &= E(a + b\xi - a - bE\xi)^2 = \\ &= b^2 E(\xi - E\xi)^2 = b^2 \cdot V\xi, \end{aligned}$$

a and b are constants. In particular,

$$Va = 0, \qquad V(b\xi) = b^2 V\xi.$$

Theorem. Let ξ and η be a random variables. Then

$$\begin{aligned} V(\xi + \eta) &= E((\xi - E\xi) + (\eta - E\eta))^2 = E[(\xi - E\xi)^2 + \\ &+ 2(\xi - E\xi)(\eta - E\eta) + (\eta - E\eta)^2] = \\ &= E(\xi - E\xi)^2 + E(\eta - E\eta)^2 + \\ &+ 2E[(\xi - E\xi)(\eta - E\eta)] = \\ &= V\xi + V\eta + 2E[(\xi - E\xi)(\eta - E\eta)]. \end{aligned}$$

Write $cov(\xi, \eta) \stackrel{def}{=} E[(\xi - E\xi)(\eta - E\eta)]$. This number is called the *covariance* of ξ and η.

If $V\xi > 0$ and $V\eta > 0$, then

$$\rho(\xi, \eta) \stackrel{def}{=} \frac{cov(\xi, \eta)}{\sqrt{V\xi \cdot V\eta}}$$

is called the *correlation coefficient* of ξ and η.

It is easy to observe that if ξ and η are independent, so are $\xi - E\xi$ and $\eta - E\eta$. Consequently by property 5 of expectations (If ξ and η are independent, then $E\xi\eta = E\xi \cdot E\eta$), we have

$$cov(\xi, \eta) = E[(\xi - E\xi)(\eta - E\eta)] = E(\xi - E\xi) \cdot E(\eta - E\eta) = 0.$$

So, if ξ and η are independent, we have $cov(\xi, \eta) = 0$! The converse isn't correct

$$cov(\xi, \eta) = 0|| \not\Longrightarrow ||P(\xi = x_j, \eta = y_k) = P(\xi = x_j)P(\eta = y_k)$$

for all $x_j \in x$ and $y_k \in y$, where x and y the set of values ξ and η respectively.

Using the notation that we introduced for covariance, we have

$$\boxed{V(\xi + \eta) = V\xi + V\eta + 2cov(\xi, \eta)}$$

Corollary. If random variables ξ and η are *independent*, the variance of the sum $\xi + \eta$ is equal to the sum of the variances

$$\boxed{V(\xi + \eta) = V\xi + V\eta}$$

Remark. The last formula is still valid under weaker hypotheses than the independence of ξ and η. In fact, it is enough to suppose that ξ and η are uncorrelated i. e.

$$cov(\xi, \eta) = 0.$$

Example. If ξ is a Bernoulli random variable, taking the values 1 and 0 with probabilities p and q, then

$$\begin{aligned} V\xi = E(\xi - E\xi)^2 = E(\xi - p)^2 = \\ = (1-p)^2 P(\xi = 1) + p^2 P(\xi = 0) = \\ = (1-p)^2 \cdot p + p^2 q = q^2 p + p^2 q = qp. \end{aligned}$$

It follows that if $\xi_1, ..., \xi_n$ are independent identically distributed Bernoulli random variables, and

$$S_n = \xi_1 + ... + \xi_n \quad \text{then} \quad VS_n = nV\xi_1 = npq.$$

Example. Let ξ be a Poisson random variable

$$P(\xi = m) = \frac{\lambda^m e^{-\lambda}}{m!}$$

Then

$$\begin{aligned}V\xi = E\xi^2-(E\xi)^2 &= \sum_{m=1}^{\infty} m^2 \tfrac{\lambda^m e^{-\lambda}}{m!} - \lambda^2 = \\ &= \lambda \sum_{m=1}^{\infty} m \tfrac{\lambda^{m-1}}{(m-1)!} e^{-\lambda} - \lambda^2 = \\ &= \lambda \sum_{m=1}^{\infty} (m-1) \tfrac{\lambda^{m-1}}{(m-1)!} e^{-\lambda} + \lambda \sum_{m=1}^{\infty} \tfrac{\lambda^{m-1}}{(m-1)!} e^{-\lambda} - \lambda^2 = \\ &= \lambda^2 + \lambda - \lambda^2 = \lambda.\end{aligned}$$

Reminder. $E\xi = \lambda$, where $P(\xi = m) = \frac{\lambda^m e^{-\lambda}}{m!}$.

12 Limit Theorems

12.1 A Miracle or a rule on a Galton desk?

Lets imagine a desk with obstacles and sections. A particle begins its way from the top to the bottom of the desk. On the first level there is only one obstacle and the particle chooses its way randomly left or right with equal possibility. On the second level there are two obstacles. The particle meets the first or the second obstacle and the situation repeats. Finally, the particle finds its place in the sections. The gadget is known as Galton's desk. If we repeat the experience many times the particle will take up its position under bell-shaped curve $\varphi(x)$ (See Section 9). The question is: Is this a casual observation or does it represents some kind of a rule? Later it will be given exact answer by limit theorem for Bernoulli trials.

Let $\xi_1, ..., \xi_n$ be independent identically distributed random variables, with $P(\xi_j = 1) = p$, $P(\xi_j = 0) = q$, $j = 1, 2, ..., n$, $p + q = 1$. This is so called James Bernoulli trials with two outcomes (success and failure) and probability p of success. Then if $S_n = \xi_1 + ... + \xi_n$ we have $ES_n = np$, $E(S_n = np)^2 = npq$.

We set the problem of finding convenient asymptotic formulas, as $n \longrightarrow \infty$, for $P(S_n = m)$ and for their sum over the values of

m that satisfy the condition

$$|x_m| = |\frac{m - np}{\sqrt{npq}}| \leq c.$$

12.2 De Moivre's local limit theorem

Let

$$\xi_1, ..., \xi_n, ...$$

be a sequence of independent Bernoulli random variables (i. e. $P(\xi_j = 1) = p,\ P(\xi_j = 0) = q,\ j = 1, 2, ..., n, p + q = 1)$ and $S_n = \xi_1 + ... + \xi_n$. As before we write $P_n(k) = C_n^k p^k q^{n-k}$ $(0 \leq k \leq n)$.

Theorem. Let $0 < p < 1$; then

$$\begin{aligned} P_n(m) &= P(S_n = m) \\ &= \frac{1}{\sqrt{2\pi npq}} \exp\left\{-\frac{(m - np)^2}{2npq}\right\} + o\left(\frac{1}{\sqrt{npq}}\right) \end{aligned}$$

uniformly for m such that $|m - np| = O(\sqrt{npq})$.

Proof. The proof depends on Stirling's formula

$$n! = \sqrt{2\pi} n^{n+1/2} e^{-n+\Theta n}, \quad \text{where} \quad \Theta_n = O\left(\frac{1}{n}\right).$$

Let's investigate the asymptotic behavior of the binomial distribution

$$P(S_n = m) = \frac{n!}{m!(n-m)!} p^m q^{n-m}.$$

We have

$$\ln P(S_n = m) = \ln n! - \ln m! - \ln(n - m)! + m \ln p + (n - m) \ln q.$$

Take into account, that

$$m = x_m\sqrt{npq} + np = n(1-q) + x_m\sqrt{npq}$$

and

$$n - m = nq - x_m\sqrt{npq}.$$

(for brevity we shall write x instead of x_m).

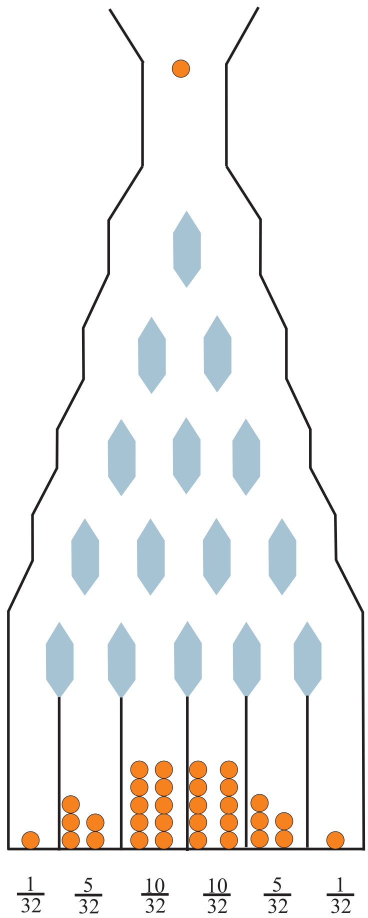


Fig. 13. The Galton's desk.

Therefore

$$
\begin{aligned}
\ln P(S_n = m) &= n \ln n + \ln \sqrt{2\pi n} - n + 0\left(\frac{1}{n}\right) - \\
&- \quad m \ln m - \ln \sqrt{2\pi m} + m + 0\left(\frac{1}{m}\right) - \\
&- \quad (n-m)\ln(n-m) - \ln \sqrt{2\pi(n-m)} + n - m + \\
&+ \quad 0\left(\frac{1}{n-m}\right) + m \ln p + (n-m)\ln q = n \ln n - \\
&- \quad m \ln m - (n-m)\ln(n-m) + \frac{1}{2}\ln \frac{2\pi n}{2\pi(n-m)2\pi} + \\
&+ \quad \underbrace{0\left(\frac{1}{n}\right) + 0\left(\frac{1}{m}\right) + 0\left(\frac{1}{n-m}\right)}_{R_n} + m \ln p + \\
&+ \quad (n-m)\ln q = -(np + x\sqrt{npq})\ln\left(1 + x\frac{\sqrt{npq}}{np}\right) - \\
&- \quad (nq - x\sqrt{npq})\ln\Big(1 - \frac{x\sqrt{npq}}{nq}\Big) + \ln\frac{1}{\sqrt{2\pi}} + \\
&+ \quad \frac{1}{2}\ln\frac{n}{m(n-m)} + R_n = \ln\frac{1}{\sqrt{2\pi}} + \frac{1}{2}\ln\frac{n}{(n-m)m} - \\
&- \quad (np + x\sqrt{npq})\left(\frac{xq}{\sqrt{npq}} - \frac{x^2q^2}{2npq} + 0\left(\frac{1}{(npq)^{3/2}}\right)\right) - \\
&- \quad (nq - x\sqrt{npq})\left(-\frac{xp}{\sqrt{npq}} - \frac{x^2p^2}{2npq} + 0\left(\frac{1}{(npq)^{3/2}}\right)\right) + \\
&+ \quad R_n = \ln\frac{1}{\sqrt{2\pi}} + \frac{1}{2}\ln\frac{n}{(n-m)m} - \frac{x^2}{2}q + \\
&+ \quad \frac{x^3q^2}{2\sqrt{npq}} - \frac{x^2p}{2} + \frac{x^3p^2}{2\sqrt{npq}} + 0\left(\frac{1}{n^{3/2}}\right) + 0\left(\frac{1}{n}\right) = \\
&= \quad \ln\frac{1}{\sqrt{2\pi}} + \frac{1}{2}\ln\frac{n}{(n-m)m} - \frac{x^2}{2} + 0\left(\frac{1}{\sqrt{n}}\right).
\end{aligned}
$$

This completes the proof.

Corollary. The conclusion of the local limit theorem can be put in the following equivalent form: For all $x \in R'$ such that $x = O(\sqrt{npq})$, and for $m = np + x\sqrt{npq}$ an integer from the set $\{0, 1, ..., n\}$

$$P_n(m) = P_n(np + x_m\sqrt{npq}) \sim \frac{1}{\sqrt{2\pi npq}} e^{-x^2/m^2},$$

i. e. as $n \to \infty$

$$P(S_n = m) = \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi npq}} \left[1 + \frac{(q-p)(x^3 - 3x)}{6\sqrt{npq}}\right] + \Delta,$$

where

$$|\Delta| < \frac{0,15 + 0,25|p-q|}{(npq)^{3/2}} + e^{-\frac{3}{2}\sqrt{npq}}$$

$$\sup_{m:\,|x_m| \le \Psi(n)} \left| \frac{P_n(np + x_m\sqrt{npq})}{e^{-x^2/2}/\sqrt{2\pi npq}} - 1 \right| \to 0$$

where $\Psi(n) = O(\sqrt{npq})$.

We can reformulate these results in probabilistic language in the following way:

$$P(S_n = k) \sim \frac{1}{\sqrt{2\pi npq}} e^{-(k-np)^2/2npq}, \qquad |k - np| = O(\sqrt{npq}),$$

$$P\left(\frac{S_n - np}{\sqrt{npq}} = x\right) \sim \frac{1}{\sqrt{2\pi npq}} e^{-x^2/2}, \qquad x = O(\sqrt{npq}).$$

If we put

$$t_k = (k - np)/\sqrt{npq} \qquad \text{and} \qquad \triangle t_k = t_{k+1} - t_k = 1/\sqrt{npq},$$

the preceding formula assumes the form

$$P\left\{\frac{S_n - np}{\sqrt{npq}} = t_k\right\} \sim \frac{\triangle t_k}{\sqrt{2\pi}} e^{-t_k^2/2}, \qquad t_k = O(\sqrt{npq}).$$

It is clear that $\triangle t_k = 1/\sqrt{npq} \to 0$, as $n \to \infty$ and the set of points $\{t_k\}$ at it were "fills" the real line. It is natural to expect that the last formula can be used to obtain the integral formula

$$P\left\{a < \frac{S_n - np}{\sqrt{npq}} \le b\right\} \sim \int\limits_a^b e^{-x^2/2} dx.$$

Let us now give a precise state

12.3 De Moivre-Laplace integral theorem

Let $0 \le p < 1, \qquad P_n(k) = C_n^k p^k q^{n-k}$. Then

$$\sup_{-\infty \le a < b \le \infty} \left| P\left(a < \frac{S_n - ES_n}{\sqrt{VS_n}} \le b\right) - \frac{1}{\sqrt{2\pi}} \int\limits_a^b e^{-\frac{x^2}{2}} dx \right| \to 0$$

$n \to \infty$. It follows at once from this formula that

$$\left| P(A < S_n \le B) - \left[\Phi\left(\frac{B - np}{\sqrt{npq}}\right) - \Phi\left(\frac{A - np}{\sqrt{npq}}\right)\right] \right| \to 0;$$

as $n \to \infty$ where

$$\Phi(x, a, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \int\limits_{-\infty}^{x} \exp\left\{-\frac{(t-a)^2}{2\sigma^2}\right\} dt;$$

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{x} e^{-\frac{u^2}{2}} du \text{ — the normal distribution function.}$$

$$\Phi(x) = \frac{1}{2} + \frac{1}{2}\left(\frac{2}{\sqrt{\pi}} \int\limits_0^x e^{-t^2} dt\right) = \frac{1}{2} + \frac{1}{2}\mathrm{erf}\left(\frac{x}{\sqrt{2}}\right).$$

Example. A true die is tossed 12000 times. We ask for the probability P that the number of G's lies in the interval $(1800, 2100)$. The required probability is

$$P_n(m) \quad = \sum_{1800 < k \le 2100} C_{12000}^k \left(\tfrac{1}{6}\right)^{1/2} \cdot \left(\tfrac{5}{6}\right)^{12000-k}.$$

An exact calculation of this sum would obviously be rather different. However, if we use the integral theorem we find that the probability P in question is ($n = 12000, p = 1/6$, $a = 1800, b = 2100$)

$$\Phi\left(\frac{2100-2000}{\sqrt{12000\cdot\frac{1}{6}\cdot\frac{5}{6}}}\right) - \Phi\left(\frac{1800-2000}{\sqrt{12000\cdot\frac{1}{6}\cdot\frac{5}{6}}}\right)$$

$$= \Phi(\sqrt{6}) - \Phi(-2\sqrt{6}) = \Phi(2{,}449) - \Phi(-4{,}889) = 0{,}992.$$

Where the values of $\Phi(2{,}449)$ and $\Phi(-4{,}898)$ were taken from tables of $\Phi(x)$ (this is normal distribution function).

It should be noted, that

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{x} e^{-t^2/2} dt = \frac{1}{2} + \frac{1}{\sqrt{2\pi}} \int\limits_{0}^{x} e^{-t^2/2} dt.$$

$$\frac{1}{\sqrt{2\pi}} \int\limits_{-x}^{x} e^{-t^2/2} dt = 1 - \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{-x} - \frac{1}{2\pi} \int\limits_{x}^{\infty} =$$

$$= \Phi(x) - \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{-x} = 2\Phi(x) - 1.$$

In some tables it is possible to meet the function

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int\limits_{0}^{x} e^{-u^2} du, \quad (x, \infty)$$

this is an error function (erf). We find that

$$\frac{1}{\sqrt{2\pi}} \int\limits_{0}^{x} e^{-t^2/2} dt = \frac{1}{2} \cdot \frac{2}{\sqrt{\pi}} \int\limits_{0}^{\frac{x}{\sqrt{2}}} e^{-t^2} dt = \frac{1}{2} \operatorname{erf}\left(\frac{x}{\sqrt{2}}\right).$$

$\Phi(x) = \frac{1}{2} + \frac{1}{2}\mathrm{erf}(\frac{x}{\sqrt{2}})$.

It is natural to ask how rapid the approach to zero is in the Moivre-Laplace Integral Theorem as $n \to \infty$. We quote a result in this direction

$$\sup_x |F_n(x) - \Phi(x)| \leq \frac{p^2 + q^2}{\sqrt{npq}},$$

where

$$F_n(x) = P\left(\frac{S_n - np}{\sqrt{npq}} \leq x\right)$$

It is important to recognize that the order of the estimate $1/\sqrt{npq}$ cannot be improved.

In this connection we note that if we change the approximation in the following way:

$$P(A < S_n \leq B) - \left[\Phi\left(\frac{B - np + 1/2}{\sqrt{npq}}\right) - \Phi\left(\frac{A - np + 1/2}{\sqrt{npq}}\right)\right] \tag{1}$$

we can get a somewhat better approximation than the approximation by De Moivre- Laplace Integral Theorem (MLIT).

A	B	Exact value $\sum C_n^m p^m q^{n-m}$	Normal approximation by MLIT	Multiple approximation by (1)
		n=100 p=0,5		
40	60	0,9648	0,9545	0,9643
45	55	0,7287	0,6827	0,7287
55	65	0,1832	0,1573	0,1831
		n=300 p=0,5		
135	165	0,9267	0,9167	0,9265
140	160	0,7747	0,7518	0,7747
160	180	0,1361	0,1238	0,1361

Finally we should remark that many of the fundamental results in probability theory are formulated as limit theorems. De Moivre-Laplace theorem was formulated as a limit theorem, which can fairly be called the origin of a genuine theory of probability and, in particular, which led the way to numerous investigations that clarified the conditions for validity of the central limit theorem. The De Moivre-Laplace theorem is the progenitor of the central limit theorem.

Central Limit Theorem (Lindeberg (1922), Levy (1925)).

Let $\xi_1, \xi_2, \dots$ be a sequence of independent identically distributed random variables with $E\xi_1^2 < \infty$ and $S_n = \xi_1 + \dots \xi_n$. Then as $n \to \infty$

$$P\left(\frac{S_n - ES_n}{\sqrt{DS_n}} \leq x\right) \to \Phi(x).$$

13 The Law of Large Numbers

Let us consider a triple $(\Omega, \mathcal{A}, P)$ with

$$\Omega = \{\omega : \ \omega = (a_1, ..., a_n), \qquad a_j = (0,1)\}$$

$\mathcal{A}$ is an algebra of subsets of Ω

$$p(\omega) = p^{\sum a_j} q^{n - \sum a_j}, \qquad p + q = 1.$$

This triple called a probabilistic model of independent experiments with two outcomes, or a *Bernoulli scheme*.

In the following part we study some limiting properties for Bernoulli trials. These are expressed in terms of random variables and of the probabilities of events connecting them.

We introduce random variables $\xi_1, ..., \xi_n$ by taking $\xi_j(\omega) = a_j \quad j = 1, 2, ..., n$ where $\omega = (a_1, ..., a_n)$. As we saw above, the Bernoulli variables $\xi_j = \xi_j(\omega)$ are independent and identically distributed

$$P(\xi_j = 1) = p, \quad P(\xi_j = 0) = q, \quad j = 1, ..., n.$$

It is natural to think of ξ_j as describing the result of an experiment at the j-th stage (or at a time j).

Let us put $S_n = S_n(\omega)$ and $S_k = \xi_1 + ... + \xi_n$. As we found above, $ES_n = np$ and consequently $E\frac{S_n}{n} = p$.

In other words, the mean value of the frequency of "success", i. e. $\frac{S_n}{n}$ coincides with the probability p of success. Hence we are led to ask how much the frequency $\frac{S_n}{n}$ of success differs from its probability p.

First of all it should be noted that we cannot expect that for sufficiently small $\varepsilon > 0$ and for sufficiently large n, the deviation $\frac{S_n}{n}$ from p is less than ε for all ω, $(S_n = S_n(\omega))$ i. e. that

$$\left|\frac{S_n}{n} - p\right| \leq \varepsilon, \quad \text{for all} \quad \omega \in \Omega. \tag{2}$$

In fact, when $0 < p < 1$,

$$P\left\{\frac{S_n}{n} = 1\right\} = P(\xi_1 = 1,\ \xi_2 = 1, ...,\ \xi_n = 1) = p^n,$$

$$P\left(\frac{S_n}{n} = 0\right) = P(\xi_1 = 0, ..., \xi_n = 0) = q^n,$$

hence it follows that (2) is not satisfied for sufficiently small $\varepsilon > 0$.

We observe, however, that when n is large the probabilities of the events $\{S_n/n = 1\}$ and $\{S_n/n = 0\}$ are small. It is therefore natural to expect that the total probability of the events for which

$$|[S_n(\omega)/n] - p| > \varepsilon$$

will also be small when n is sufficiently large.

We shall accordingly try to estimate the probability of the event $\omega :\ |[S_n(\omega)/n] - p| > \varepsilon$. For this purpose we need the following inequality.

Chebyshev's inequality. **(Chebyshev Pafnuti (1821–1894) — Russian mathematicians).** Let $(\Omega, \mathcal{A}, \mathcal{P})$ be a probability space and $\xi = \xi(\omega)$ be a nonnegative random variable $(\xi \geq 0)$. Then, *Markov's inequality* **(Markov Andrey (1856–1922) — Russian mathematicians)**

$$E\xi = \sum_j x_j p_j \geq \sum_{j:xj>\varepsilon} = \varepsilon \sum_{j:xj>\varepsilon} p_j = \varepsilon p(\xi > \varepsilon).$$

$$P(\xi \geq \varepsilon) \leq E\xi/\varepsilon \quad \text{for all} \quad \varepsilon > 0. \qquad (3)$$

Proof. We notice that

$$\xi = \xi I(\xi \geq \varepsilon) + \xi I(\xi < \varepsilon) > \xi I(\xi \geq \varepsilon) \geq \varepsilon I(\xi \geq \varepsilon),$$

where $I(A)$ is the indicator of A.

Then, by the properties of expectation,

$$E\xi \geq \varepsilon EI(\xi \geq \varepsilon) = \varepsilon P(\xi \geq \varepsilon), \text{which establishes (3).}$$

Corollary. If ξ is any random variable, we have for all $\varepsilon > 0$

$$\begin{gathered} P\{|\xi| \geq \varepsilon\} \leq E|\xi|/\varepsilon, \\ P\{|\xi| \geq \varepsilon\} = P(\xi^2 \geq \varepsilon^2\} \leq \tfrac{E\xi^2}{\varepsilon^2} \\ P\{|\xi - E\xi| \geq \varepsilon\} \leq V\xi/\varepsilon^2. \end{gathered}$$

In the last of these inequalities, take $\xi = S_n/n$ and take in account that if $\xi_1, ..., \xi_n$ are i.i.d. Bernoulli random variable and $S_n = \xi_1 + ... + \xi_n$ then $VS_n = npq$, we obtain

$$P\left(\left|\frac{S_n}{n} - p\right| \geq \varepsilon\right) \leq \frac{V(S_n/n)}{\varepsilon^2} = \frac{VS_n}{n^2\varepsilon^2} = \frac{npq}{n^2\varepsilon^2} = \frac{pq}{n\varepsilon^2}.$$

Therefore

$$P\left\{\left|\frac{S_n}{n} - p\right| \geq \varepsilon\right\} \leq \frac{pq}{n\varepsilon^2} \leq \frac{1}{4n\varepsilon^2}.$$

So we have

Theorem (J. Bernoulli 1713). The probability that the frequency $\frac{S_n}{n}$ differs from its mean ($E\frac{S_n}{n} = p$) value p by a quantity of modulus at least equal to ε tends to zero as $n \to \infty$, however small $\varepsilon > 0$ is chosen.

Reminder.

$$I(A) = \begin{cases} 1 & \omega \in A \\ 0 & \omega \bar{\in} A \end{cases}$$

$$EI(A) = 1 \cdot P(A) + 0 \cdot P(\bar{A}) = P(A).$$

Appendix:

Order of magnitude as $x \longrightarrow \infty$

Here we introduce the little "o" [ou] and big "O" [ou] notation invented by number theorists a hundred years ago and now commonplace in mathematical analysis and computer science.

Definition. A function f is of smaller order than g as $x \to \infty$ if

$$\lim_{x\to\infty} \frac{f(x)}{g(x)} = 0.$$

We indicate this by writing $f = o(g)$; ("f is little-oh of g"). Notice that saying $f = o(g)$ as $x \to \infty$ is another way of saying that f grows slower than g as $x \to \infty$.

Example. $\ln x = o(x)$ as $x \to \infty$ because

$$\lim_{x\to\infty} \frac{\ln x}{x} = 0.$$

$x^2 = o(x^3 + 1)$ as $x \to \infty$ because

$$\lim_{x\to\infty} \frac{x^2}{x^3+1} = 0.$$

Definition. A function f is of at most the order of g as $x \to \infty$ if there is a positive integer M for which $\frac{f(x)}{g(x)} \leq M$, for x sufficiently large. We indicate this by writing $f = O(g)$ ("f is big-oh of g").

Example. $x + \sin x = O(x)$ as $x \to \infty$ because $\frac{x+\sin x}{x} \leq 2$ for x sufficiently large and $M = 2$.

Example. $e^x + x^2 = O(e^x)$ as $x \to \infty$ and $M = 2$ because $ax + b = O(x)$ as $x \to \infty - M = a + 1$.

If you look at the definitions again, you will see that $f = o(g)|| \Longrightarrow ||f = O(g)$.

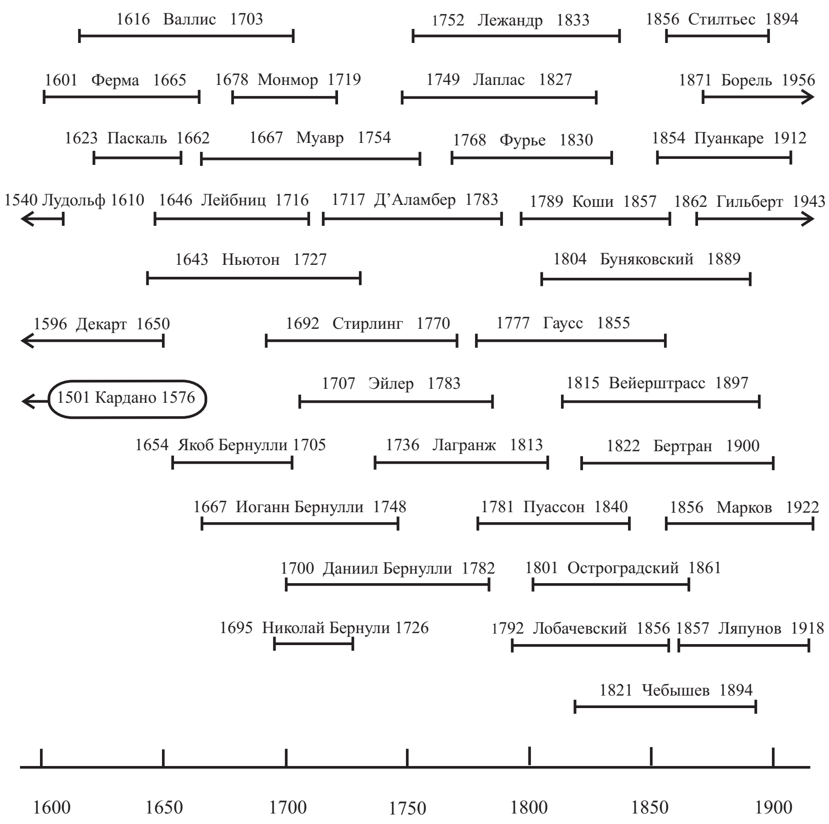
Definition. A function f is asymptotically equal to $g(x)$ if

$$\lim_{x\to\infty} \frac{f(x)}{g(x)} = 1.$$

We write $f(x) \sim g(x)$.

Example. $\frac{x^2}{x+\log x} \sim x$ as $x \to \infty$. $\sin x \sim x$ as $x \to 0$. Notice that O(1) signifies "any bounded function"; and o(1) "any function tending to zero".

Хронологическая таблица



References

1. Bolshev L.N., Smirnov N.V. Tables of mathematical statistic. M. : Nauka, 1983. (Russian).
2. Breiman L. Probability. 2d printing. Philadelphia : SIAM, 1993.
3. Cramer H. Mathematical methods of statistics. Princeton, N.J. : Princeton univ. press, 1957.
4. Dixon W., Massey F., jr. Introduction to statistical analysis. 2d ed. New York etc. : McGraw-Hill, 1957.
5. Fastovec N.O. Elementary theory of probability and mathematical statistics. M., 1991. (Russian).
6. Feller W. An Introduction to probability theory and its applications : In 2 vol. 2d ed. New York etc. : Wiley, cop. 1957–1971.
7. Handbook of mathematical functions / Ed. by M. Abramovitz a. I.A. Stegun. M. : Nauka, 1979. (Russian).
8. Kreyszig E. Advanced engineering mathematics. New York : Wiley, 1988.
9. Loeve M. Probability theory. New York etc. : Springer, 1977–1978.
10. Mosteller F., Rourke R.E.K., Thomas G.B., jr. Probability : A first course. Reading, Mass.; London : Addison Wesley, 1961.
11. Prohorov Yu.V., Rozanov Yu.A. Probability theory. New York etc. : Springer, 1969.
12. Rotar V.I. Theory of probability. M. : Vissh. shk., 1992. (Russian).
13. Saporta G. Probabilités analyse des données et statistique. Paris : Technip, 1990.
14. Shiryaev A. Probability. New York etc. : Springer, cop. 1996.

Contents

Учебное пособие

Гамкрелидзе Николай Георгиевич

Элементарная теория вероятностей

Техническая обработка и подготовка материала:
Нина Гамкрелидзе

Оригинал-макет подготовлен с использованием издательской системы LaTeX

Подписано к использованию: 10.04.2019
Объем издания 0,367 Мб; Тираж: 50 экз.
Комплектация издания: 1 CD-ROM;
Запись на физический носитель: Белоусов А.В.,
belousov.a@gubkin.ru
119991, г. Москва, Ленинскй проспект, д. 65, корпус 1
РГУ нефти и газа (НИУ) имени И.М. Губкина,
учебно-методическое управление (900)