Министерство науки и высшего образования Российской Федерации

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКИЙ ГОСУДАРСТВЕННЫЙ ГИДРОМЕТЕОРОЛОГИЧЕСКИЙ УНИВЕРСИТЕТ

С.Н. Фадеев, И.В. Зайцева

ТЕОРИЯ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Санкт-Петербург
РГГМУ
2021

УДК 519.2
ББК 22.171
Ф15

Рецензент: Нестеренко Г.Н., доцент кафедры информатики и информационных таможенных технологий Санкт-Петербургского им. В.Б. Бобкова филиала Российской таможенной академии, канд. техн. наук, доцент

Фадеев С.Н., Зайцева И.В.
Ф15 Теория вероятностей и математическая статистика / С.Н. Фадеев, И.В. Зайцева. – Санкт-Петербург : РГГМУ, 2021. – 92 с.

Излагаются основные вопросы дисциплины «Теория вероятностей и математическая статистика». Подробно рассматриваются решения типовых задач; в конце каждого раздела приведены задачи для самостоятельного решения с ответами. Пособие содержит необходимый справочный материал и список литературы.

Рекомендовано для самостоятельной работы студентов, а также для использования в качестве дополнительного материала при проведении преподавателем лекционных и практических занятий.

ISBN 978-5-86813-536-1

Оглавление

Глава 1. Основные понятия теории вероятностей.

1.1. Случайное событие, его вероятность

Элементы комбинаторики. Прежде чем формулировать основные определения элементарной теории вероятностей, дадим здесь несколько понятий из комбинаторики – раздела математики, изучающего перестановки и комбинации предметов*.

1. Пусть имеется n предметов. Сколько имеется *перестановок* этих предметов. То есть, сколькими способами можно переупорядочить эти предметы, иначе, сколькими способами можно распределить n предметов по n местам.

Число перестановок равно произведению натуральных чисел от единицы до данного:

$$n! = 1 \cdot 2 \cdot 3 \cdot \ldots \cdot n. \qquad (1.1)$$

Сама операция, обозначаемая в математике знаком «!», называется факториал.

Пример. Пусть за первыми тремя столами в аудитории сидят шесть студентов (два студента за каждым столом). Сколькими способами можно разместить студентов по шести местам?

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

Имеет место следующее очевидное равенство $(n+1)! = n!(n+1)$.

Следует иметь в виду, что 0!=1.

2. Имеется n предметов. Сколькими способами из n предметов можно выбрать m предметов. Можно показать, что число способов равно:

$$C_n^m = \frac{n!}{m!(n-m)!}. \qquad (1.2)$$

Данное выражение называется *числом сочетаний из n по m,* а также *биномиальным коэффициентом.*

Имеют место следующие равенства, которые следуют из самой постановки задачи о числе сочетаний

$$C_n^1 = n,$$

$$C_n^{n-m} = C_n^m.$$

3. Имеется n предметов. Сколькими способами из n предметов можно выбрать m предметов, но с учетом порядка, в котором отбираются предметы. Число способов определяется выражением, которое называется *число размещений из n по m*:

$$A_n^m = \frac{n!}{(n-m)!}. \qquad (1.3)$$

Еще раз подчеркнем отличие числа сочетаний от числа размещений: «сочетание» – порядок в котором располагаются m отобранных предметов не важен, «размещение» – порядок отбора важен.

Примеры.

1) Имеется десять точек, никакие три из которых не лежат на одной прямой. Сколько различных прямых можно провести через эти точки.

---

* Здесь слово «предмет» носит условный характер, оно также может означать, например, животных или людей.

Решение. Так как две точки определяют прямую линию единственным образом, число прямых равно числу различных пар точек, то есть числу способов, сколькими из десяти точек можно выбрать две. Таким образом, число прямых равно

$$C_{10}^{2}=\frac{10!}{2!(10-2)!}=\frac{1\cdot 2\cdot 3\cdot\ldots\cdot 8\cdot 9\cdot 10}{2\cdot 1\cdot 2\cdot 3\cdot\ldots\cdot 8}=\frac{9\cdot 10}{2}=45.$$

2) В группе 15 студентов. Для участия в студенческой конференции необходимо отобрать трех студентов. Сколькими способами это можно сделать?

Ответ: $C_{15}^{3}=\frac{15!}{3!12!}=\frac{13\cdot 14\cdot 15}{2\cdot 3}=455$.

3) Две последние цифры телефонного номера, записанного на бумаге, оказались затертыми, но известно, что цифры не совпадают. Какое *максимальное* количество номеров необходимо перебрать, чтобы получить правильный номер?

Ответ: $A_{10}^{2}=\frac{10!}{8!}=90$. Здесь, в отличие от предыдущих примеров важен порядок, в котором отбираются «предметы» – цифры, так как перестановка цифр дает другое число. Разумеется, данную задачу можно решить и по-другому. Всего комбинаций двух цифр 100: 00, 01, …,99. Но следует исключить комбинации с одинаковыми цифрами, а их десять: 00, 11, 22, ..., 99. 100–10=90.

4) Расписание дня состоит из четырех занятий. Всего дисциплин семь. Сколькими способами можно составить расписание учебного дня.

Решение. Так как расписание дня зависит не только от состава дисциплин, но и порядка в котором они следуют, число способов равно

$$A_{7}^{4}=\frac{7!}{4!}=5\cdot 6\cdot 7=210.$$

Частотное и классическое определения вероятности. Пусть в некотором опыте наблюдается некоторое событие (опыт не обязательно должен быть поставлен человеком, человек может выступать в роли пассивного наблюдателя). Если результаты опыта варьируются при его повторении, говорят, что опыт имеет случайный исход. Случайным событием (или просто событием) называется всякий факт, который в опыте со случайным исходом может произойти, но может и не произойти.

Численной или количественной характеристикой того, насколько возможно (или невозможно) случайное событие является вероятность данного события.

Пусть производится n опытов (иначе, некоторый опыт повторен n *раз*). Пусть событие A наблюдалось в m_A опытах. Назовем *частотой* ν события A в n опытах отношение

$$\nu=\frac{m_A}{n}. \qquad (1.4)$$

Примеры.

1) Монета подброшена 100 раз, герб появился 47 раз, тогда частота появления герба равна 0,47.

2) Игральный кубик подброшен 20 раз. Шестерка появилась четыре раза. Частота появления шестерки равна 0,2.

Вероятностью $P(A)$ события A называется предел отношения числа опытов m_A, в которых A имело место к общему числу опытов n, когда общее число опытов неограниченно возрастает:

$$P(A) = \lim_{n\to\infty} \frac{m_A}{n}. \qquad (1.5)$$

Данное определение называют частотным определением вероятности. На интуитивном уровне частотное определение отражает связь вероятности и частоты. (подробно о характере этой связи будет сказано в п.3.1). Из этого определения следует фундаментальное свойство вероятности

$$0 \le P(A) \le 1.$$

Однако, определение (1.5) не позволяет точно вычислить вероятность и поэтому является неудовлетворительным. Произведем следующий мысленный опыт. Пусть нам надо определить вероятность появления герба при подбрасывании монеты. Будем подбрасывать монету все большее число раз, подсчитывая количество выпадений герба и соответствующую частоту:

n	m_A	ν
10	4	0,40
100	47	0,47
500	255	0,51
1000	480	0,48
...	...	...

Закономерности в последовательности возможных значений частоты ν мы не видим. В таких случаях говорят, что ν является случайной величиной. Вычислить предел случайной последовательности не представляется возможным.

С другой стороны, интуитивно мы понимаем, что вероятность появления герба при подбрасывании монеты равна 1/2. Мы исходим при этом из того, что при подбрасывании монеты есть два исхода «герб», «решка» и эти исходы равновозможны (в таких случаях говорят также, что опыт обладает симметрией возможных исходов). Отношение числа исходов, в которых данное событие имеет место к общему числу исходов и даст искомую вероятность. Таким образом, приходим к следующему определению.

Пусть все исходы данного опыта равновозможны. Пусть n – общее число исходов опыта, m – число благоприятных исходов, т.е. исходов в которых интересующее нас событие A имеет место. Тогда вероятность $P(A)$ определяется как

$$P(A) = \frac{m}{n}. \qquad (1.6)$$

Данное определение называют классическим определением вероятности. Очевидно, это определение носит ограниченный характер, так как справедливо только для равновозможных исходов. Например, если мы верим в то, что все исходы при бросании игрального кубика равновозможны, то в соответствии с (1.6) вероятность появления «6» равна 1/6. Но если у нас есть основания в этом сомневаться, применить соотношение (1.6) мы не можем.

Рассмотрим примеры применения классического определения вероятности.

1) Бросаются два игральных кубика. Найти вероятности следующих событий.

A – сумма выпавших очков четная, причем на грани хотя бы одного кубика выпадет шесть очков;

B – сумма выпавших очков нечетная, причем на грани хотя бы одного кубика выпадет единица;

C – сумма выпавших очков равна 8;

D – произведение выпавших очков равно 8;

E – сумма выпавших очков больше чем их произведение.

Полное количество исходов, т.е. знаменатель в (1.6) , равно $6 \cdot 6 = 36$.

Будем записывать рассматриваемые события как множества, элементы которых представляют собой упорядоченные пары чисел – количество очков, выпавших на первом и втором кубиках. После этого число благоприятных исходов, т.е. числитель в формуле (1.6), определяется простым подсчетом элементов множества

$$A = \{(2,6),(4,6),(6,6),(6,4),(6,2)\},\ P(A) = \frac{5}{36}.$$

$$B = \{(1,6),(1,4),(1,2),(2,1),(4,1),(6,1)\},\ P(B) = \frac{6}{36} = \frac{1}{6}.$$

$$C = \{(2,6),(3,5),(4,4),(5,3),(6,2)\},\ P(C) = \frac{5}{36}.$$

$$D = \{(2,4),(4,2)\},\ P(D) = \frac{1}{18}.$$

Для того, чтобы сумма выпавших очков была больше, чем их произведение, по крайней мере, на одном кубике должна выпасть единица. Число соответствующих исходов равно 11. Искомая вероятность $P(E) = 11/36$.

2) Из пяти карточек разрезной азбуки составлено слово «книга». Ребенок не умеющий читать рассыпал карточки, затем сложил их в случайном порядке. Какова вероятность, что у него снова получится слово «книга».

Здесь предполагается, что все исходы равновозможны. Общее число исходов равно числу перестановок пяти карточек, $n = 5! = 120$. Благоприятный исход один. Искомая вероятность $P = 1/120$.

3) Из колоды карт в 36 листов вынимают две карты. Какое событие является более вероятным: A – обе карты одного цвета, B – карты разных цветов?

Общее число исходов равно числу способов сколькими из 36 карт можно выбрать две, т.е. C_{36}^2. Так как число карт одного цвета равно 18 несложно получить

$$P(A) = \frac{2C_{18}^2}{C_{36}^2} = \frac{2\frac{18!}{2!\cdot 16!}}{\frac{36!}{2!\cdot 34!}} = \frac{2\frac{17\cdot 18}{2!}}{\frac{35\cdot 36}{2}} = \frac{17\cdot 18}{35\cdot 18} = \frac{17}{35},$$

$$P(B) = \frac{C_{18}^1 C_{18}^1}{C_{36}^2} = \frac{18\cdot 18}{\frac{36!}{2!\cdot 34!}} = \frac{18\cdot 18}{\frac{35\cdot 36}{2}} = \frac{18}{35}.$$

4) Имеется 12 деталей, из которых 5 являются бракованными. Наугад отбирается 4 детали. Какова вероятность того, что все четыре окажутся бракованными? Только одна из четырех окажется бракованной?

Общее число исходов равно C_{12}^4. Число способов сколькими можно выбрать 4 бракованных из пяти равно $C_5^4 = 5$. Поэтому вероятность, что все четыре детали окажутся бракованными, равна

$$P = \frac{C_5^4}{C_{12}^4} = \frac{5}{12!/(4!\cdot 8!)} = \frac{5}{9\cdot 5\cdot 11} = \frac{1}{99}.$$

Одну бракованную деталь из пяти бракованных можно выбрать пятью способами: $C_5^1 = 5$. Остальные три должны быть небракованными. Их можно выбрать C_7^3 способами. Полное число исходов с одной бракованной и тремя небракованными деталями равно произведению $C_5^1 C_7^3$. Таким образом, искомая вероятность равна

$$P = \frac{C_5^1 C_7^3}{C_{12}^4} = \frac{5 \cdot \frac{7}{3! \cdot 4!}}{\frac{12!}{4! \cdot 8!}} = \frac{5 \cdot 5 \cdot 7}{9 \cdot 5 \cdot 11} = \frac{35}{99}.$$

Указание. Полезно найти решение более общей задачи. Пусть имеется N предметов из которых M предметов обладают некоторым определенным качеством. Отбирается n предметов. Какова вероятность, что среди отобранных предметов ровно m предметов обладают данным качеством?

Общая формула

$$P = \frac{C_M^m \cdot C_{N-M}^{n-m}}{C_N^n}. \tag{1.7}$$

«Геометрическая» вероятность. Классическое определение вероятности (1.6) предполагает, что количество исходов опыта конечно. Рассмотрим случай, когда число исходов бесконечно, а сами исходы являются элементами, вообще говоря, несчетного множества.

Именно, требуется найти вероятность события A в том случае, когда результат испытания определяется случайным положением точек в некоторой области, причем любые положения точек в этой области являются равновозможными.

В этом случае вероятность события определяется как

$$P(\mathrm{A}) = \frac{\Omega_m}{\Omega_n}, \tag{1.8}$$

где Ω_n – мера (размер) всей области, Ω_m – мера области, попадание в которую благоприятствует событию А.

Замечание. Единицы измерения областей могут быть самыми различными, в зависимости от смысла задачи это могут быть: длина, площадь, объем, время…

Примеры

1) Расстояние между пунктами *A*, *B* равно 12 км. Между пунктами натянута телефонная линия. Найти вероятность того, что обрыв произойдет не далее чем в одном километре от одного из пунктов. Предполагается, что

- длина проволоки равна расстоянию между пунктами,
- вероятность обрыва на некотором участке не зависит от расположения этого участка.

Искомая вероятность равна отношению соответствующих длин

$$P = \frac{2}{12} = \frac{1}{6}.$$

2) В круг вписан равносторонний треугольник. Найти вероятности следующих событий. Событие *A* – точка, поставленная произвольным образом внутри круга, окажется и внутри треугольника, событие *B* – точка окажется вне треугольника.

Искомые вероятности равны отношению соответствующих площадей. Обозначив за R радиус круга, получим

$$P(A)=\frac{\frac{3\sqrt{3}}{4}R^2}{\pi R^2}=\frac{3\sqrt{3}}{4\pi},$$

$$P(B)=\frac{\pi R^2-\frac{3\sqrt{3}}{4}R^2}{\pi R^2}=1-\frac{3\sqrt{3}}{4\pi}.$$

Основные правила теории вероятностей (аксиоматическое определение вероятности).

Все приведенные выше определения понятия носят ограниченный характер. Они не могут претендовать на роль универсального определения. Это не случайно. «Вероятность» следует отнести к числу фундаментальных, самых общих, «первоосновных» понятий. Классическое определение какого-либо понятия подразумевает, прежде всего, подведение его под понятие более широкое: «яблоко это фрукт…», «студент это учащийся…» и т.д. Для вероятности, не определяя сам этот термин обычным образом, формулируют только правила, которым она подчиняется. Эти правила не выводятся логическим путем из более общих определений, т.е. являются аксиомами, соответственно, набор таких правил называют аксиоматическим определением.

Прежде чем переходить к формулировке аксиом, уточним понятие «событие». Событие в теории вероятностей рассматривается как множество. Множество всех возможных исходов данного опыта называется пространством элементарных событий. Далее будем обозначать это множество буквой Ω. Каждый элемент $\omega\in\Omega$ называется элементарным событием.

Приведем *примеры*.

1) Бросается игральный кубик. Пространство элементарных событий – множество, элементами которого являются первые шесть натуральных чисел $\Omega=\{1,2,3,4,5,6\}$. Событие A, которое заключается в выпадении четного числа очков это множество $A=\{2,4,6\}$. Событие B – выпадение нечетного числа очков, множество $B=\{1,3,5\}$. Событие M – выпадение максимально возможного числа очков, это множество, состоящее из одного элемента: $M=\{6\}$.

2) Бросаются два игральных кубика. Множество Ω состоит из упорядоченных пар чисел от 1 до 6:

$$\Omega=\{(1,1),(1,2),\ldots,(6,6)\}.$$

Событие A – сумма выпавших очков четная, причем на грани хотя бы одного кубика выпадет шесть очков это множество

$$A=\{(2,6),(4,6),(6,6),(6,4),(6,2)\}.$$

3) Из пяти карточек разрезной азбуки случайным образом составляются слова. Пространство элементарных событий – множество из $5!=120$ элементов, каждый элемент которого представляет собой определенную перестановку пяти предметов – карточек с буквами.

4) Стрелок стреляет по мишени, которая представляет собой круг радиуса R. Максимальному числу очков – «десятке» соответствует внутренний круг радиуса r.

Введем декартову прямоугольную систему координат xOy с центром, совпадающем с центром мишени. Множество Ω – множество всех точек (x, y) плоскости.

Событие A – стрелок попал в мишень это множество точек

$$A = \{(x, y) : x^2 + y^2 \le R^2\}$$

Событие B – стрелок попал в «десятку»:

$$B = \{(x, y) : x^2 + y^2 \le r^2\}.$$

Напомним, что *объединением* или *суммой* множеств A и B называется множество, состоящее из всех элементов каждый из которых принадлежит хотя бы одному из данных множеств. Обозначение $A \cup B$ или $A + B$, причем далее мы будем пользоваться именно вторым обозначением.

Сумму событий можно трактовать как объединение соответствующих множеств. Но мы дадим более конкретное определение.

Суммой событий A и B ($A + B$) называется событие, состоящее в появлении хотя бы одного из двух данных событий.

Пересечением или *произведением* множеств A и B называется множество каждый элемент которого принадлежит как A так и B. Обозначения $A \cap B$ или $AB \equiv A \cdot B$. Далее будет пользоваться вторым обозначением.

Произведению событие соответствует пересечение соответствующих множеств.

Более конкретно: произведением событий A и B (AB) называется событие состоящее в совместном появлении данных событий (как A так и B).

Говорят, что события $A_1, A_2, \ldots, A_n$, как исходы некоторого опыта, образуют полную группу событий, если в результате опыта неизбежно должно появиться хотя бы одно из них.

Например, при бросании игрального кубика события A – выпадение четного числа очков, B – выпадение нечетного числа очков, M – выпадение максимально возможного числа очков, образуют полную группу событий. Полную группу событий образуют и события A, B.

Для полной группы событий имеет место очевидное равенство

$$A_1 + A_2 + \ldots + A_n = \Omega.$$

Если события A, B не могут появиться вместе (говорят, что они являются несовместными), то, очевидно, $AB = \varnothing$.

Если A, B несовместны, то их суммой является событие, состоящее в появлении *или A или B.*

Определения суммы и произведения событий, очевидным образом обобщаются на более чем два события.

Для данного события A *событие,* обозначаемое как $\overline{A}$, называется *противоположным* A, если оно состоит в непоявлении A.

События A и $\overline{A}$ образуют полную группу событий и несовместны, т.е. имеют место следующие свойства:

$$A + \overline{A} = \Omega,$$

$$A \cdot \overline{A} = \varnothing.$$

Теперь, основываясь на приведенной трактовке события, переходим к формулировке основных правил (аксиом) теории вероятностей.

1) Вероятность случайного события A это число $P(A)$, которое сопоставляется данному событию.

2) $0 \le P(A) \le 1$, причем $P(\Omega) = 1$. Событие, вероятность которого равна нулю (невозможное событие) соответствует пустому множеству, то есть $P(\varnothing) = 0$.

3) Пусть $A_1, A_2, \ldots, A_n$ попарно несовместные события ($A_i A_j = \varnothing, i \ne j$). Вероятность суммы несовместных событий равна сумме их вероятностей:

$$P(A_1 + A_2 + \ldots + A_n) = P(A_1) + P(A_2) + \ldots + P(A_n).$$

4) Пусть имеется бесконечное, но *счетное* множество попарно несовместных событий $A_1, A_2, A_3 \ldots$ Тогда имеет место расширенная аксиома сложения:

$$P(A_1 + A_2 + A_3 + \ldots) = P(A_1) + P(A_2) + P(A_3) + \ldots.$$

5) Правило умножения вероятностей*:

$$P(AB) = P(A)P(B \mid A),$$

где $P(B \mid A)$ – условная вероятность: вероятность события *B при условии*, что событие *A* произошло. Так как никакого хронологического упорядочивания здесь нет, то правило умножения можно переписать и как

$$P(AB) = P(B)P(A \mid B).$$

В том случае, когда события независимы (появление одного из них никак не влияет на появление или непоявление другого) понятие условной вероятности теряет смысл и правило умножения приобретает наиболее простой вид

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

Следствия из правил.

1) Так как некоторое событие *A* и противоположное событие $\overline{A}$ образуют полную группу имеем

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

Так как, события А и $\overline{A}$ несовместны применяя правило 3) получаем

$$P(A) + P(\overline{A}) = 1. \qquad (1.9)$$

Несмотря на свою простоту, данное правило широко используется при вычислении вероятности. Часто непосредственное вычисление вероятности некоторого события представляет весьма трудоемкую задачу. Но вычисление вероятности противоположного события выполнить при этом несложно.

2) Рассмотрим два события *A, B не обязательно несовместных.* Как любое множество, событие можно условно представить в виде множества точек – круга (круги, используемые для изображения множеств, называются кругами Эйлера). Множества *A, B* показаны на рис. 1.

*Данное правило не включается в число аксиом, но является следствием определения условной вероятности. См., например, Колмогоров А.Н. Основные понятия теории вероятностей. 2 изд. - М.: Наука, 1974, 120 с.

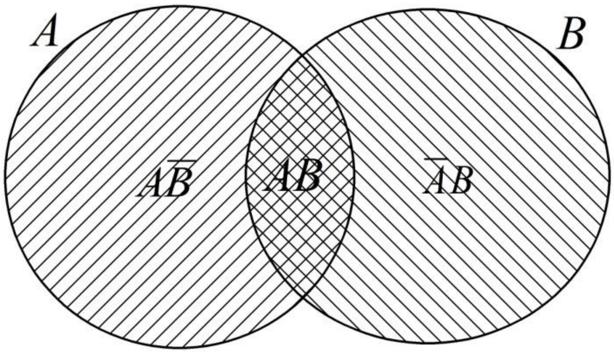


Рис.1.1

Из рисунка видим

$A + B = A\overline{B} + AB + \overline{A}B$.

Так как множества $A\overline{B}$, $AB, \overline{A}B$ не пересекаются, соответствующие события несовместны, поэтому

$$P(A+B) = P(A\overline{B}) + P(AB) + P(\overline{A}B). \qquad (1.10а)$$

Множество A можно представить как объединение двух непересекающихся множеств

$A = A\overline{B} + AB$.

Следовательно

$$P(A) = P(A\overline{B}) + P(AB). \qquad (1.10б)$$

Аналогично

$$P(B) = P(\overline{A}B) + P(AB). \qquad (1.10в)$$

Из равенств (1.10а), (1.10б), (1.10в) несложно получить выражение для вероятности суммы двух произвольных событий

$$P(A+B) = P(A) + P(B) - P(AB). \qquad (1.11)$$

Примеры.

1) Три стрелка стреляют по мишени, делая по одному выстрелу. Вероятность попадания для первого стрелка равна 0.6, для второго – 0.7, третьего – 0.85. Найти вероятности следующих событий:

A – в мишени ровно одна пробоина,

B – в мишени хотя бы одна пробоина,

C – в мишени нет пробоин.

Предполагается, что стрелки попадают в мишень независимо друг от друга.

Обозначим как S_1 событие – первый стрелок попал в мишень, S_2 – второй стрелок попал в мишень, S_3 – третий стрелок попал. Если в мишени только одна пробоина, то в мишень попал только один стрелок. Поэтому событие A можно представить как

$A = S_1\overline{S}_2\overline{S}_3 + \overline{S}_1 S_2 \overline{S}_3 + \overline{S}_1\overline{S}_2 S_3$.

Напомним, что черта над буквой означает событие противоположное данному. Все слагаемые представляют собой несовместные события. Поэтому

$$P(A) \overset{3)}{=} P(S_1\bar{S}_2\bar{S}_3) + P(\bar{S}_1 S_2\bar{S}_3) + P(\bar{S}_1\bar{S}_2 S_3) \overset{5)}{=}$$
$$= P(S_1)P(\bar{S}_2)P(\bar{S}_3) + P(\bar{S}_1)P(S_2)P(\bar{S}_3) + P(\bar{S}_1)P(\bar{S}_2)P(S_3),$$
где число над знаком равенства обозначает номер правила, которое было использовано.

Для вычисления вероятностей противоположных событий воспользуемся правилом (1.9). Тогда

$$P(A) = 0{,}6\cdot(1-0{,}7)\cdot(1-0{,}85) + (1-0{,}6)\cdot 0{,}7\cdot(1-0{,}85) + (1-0{,}6)\cdot(1-0{,}7)\cdot 0{,}85 = 0{,}171.$$

Представив события B как

$$B = \underbrace{S_1\bar{S}_2\bar{S}_3 + \bar{S}_1 S_2\bar{S}_3 + \bar{S}_1\bar{S}_2 S_3}_{\text{однапробоина}} + \underbrace{S_1 S_2\bar{S}_3 + S_1\bar{S}_2 S_3 + \bar{S}_1 S_2 S_3}_{\text{двепробоины}} + \underbrace{S_1 S_2 S_3}_{\text{три пробоинь}},$$

его вероятность можно было бы найти аналогично тому, как это было сделано для вероятности $P(A)$. Но это было бы нерационально. Поэтому сначала найдем вероятность события C. Событие C является произведением (совместным выполнением) трех независимых событий: $\bar{S}_1$, $\bar{S}_2$, $\bar{S}_3$. Поэтому

$$P(C) = P(\bar{S}_1\bar{S}_2\bar{S}_3) = P(\bar{S}_1)P(\bar{S}_2)P(\bar{S}_3) = (1-0{,}6)(1-0{,}7)(1-0{,}85) = 0{,}018.$$

«В мишени хотя бы одна пробоина» и в «в мишени нет пробоин» – являются противоположными событиями. Поэтому, воспользовавшись соотношением (1.9) получим

$$P(B) = 1 - P(C) = 0{,}982.$$

2) Из колоды карт в 36 листов вынимают две карты. Какое событие является более вероятным: A – обе карты одного цвета, B – карты разных цветов?

Данная задача уже решалась с применением формулы (1.6). Теперь решим ее с помощью правила умножения вероятностей. Пусть C событие, состоящее в том, что вытащена карта определенного, но не важно какого цвета. Очевидно, $P(C) = 1$. Вероятность того, что вторая карта будет такого же цвета можно записать как условную вероятность $P(C\,|\,C)$. Эта вероятность равна $P(C\,|\,C) = \dfrac{17}{35}$, так как при извлечении карты уменьшиться общее число карт и число карт данного цвета. Тогда

$$P(A) = P(CC) = P(C)P(C\,|\,C) = 1\cdot\frac{17}{35} = \frac{17}{35}.$$

Вероятность $P(B)$ найдем как вероятность события противоположного A:

$$P(B) = 1 - \frac{17}{35} = \frac{18}{35}.$$

3) Система подачи воды на последний этаж высотного здания состоит из двух насосов. Вероятность отказа за определенный промежуток времени для первого насоса составляет $p_1 = 0{,}05$, для второго $p_2 = 0{,}04$. Найти надежность системы (вероятность безотказной работы за данный промежуток времени), если насосы соединены

а) последовательно, в этом случае отказ хотя бы одного насоса приводит к прекращению подачи воды;

б) параллельно, подача воды прекратится в случае отказа обоих насосов.

Предполагается, что насосы ломаются независимо друг от друга.

Решение

Обозначим как A событие, состоящее в отказе первого насоса, B – второго. Таким образом $\overline{A}$, $\overline{B}$ – события, состоящие в безотказной работе первого и второго насосов соответственно.

Искомые вероятности

а) $P(\overline{A}\overline{B}) = P(\overline{A})P(\overline{B}) = (1-p_1)(1-p_2) = 0{,}95 \cdot 0{,}96 = 0{,}912$;

б) $P(\overline{A}\overline{B} + A\overline{B} + \overline{A}B) = 1 - P(AB) = 1 - p_1 p_2 = 1 - 0{,}05 \cdot 0{,}04 = 0{,}998$.

1.2. Формула полной вероятности

Пусть производится опыт со случайным исходом в результате которого может появиться некоторое событие A. Требуется найти $P(A)$. Но условия опыта реализуются случайным образом, т.е. фактически сами являются случайными событиями. Априори об условиях опыта можно сделать n взаимоисключающих гипотез H_1, H_2,..., H_n , которые образуют полную группу событий (опыт обязательно протекает в соответствии с одной из гипотез):

$$H_i H_j = \varnothing, i \neq j,$$

$$H_1 + H_2 + ... + H_n = \Omega.$$

Вероятности $P(H_1)$, $P(H_2)$,…, $P(H_n)$ реализации гипотез известны. Известны также условные вероятности $P(A|H_1)$, $P(A|H_2)$,…, $P(A|H_n)$. Здесь $P(A|H_i)$ – вероятность того, что в результате опыта появится событие A при условии, что опыт протекал в соответствии с гипотезой H_i.

Вывод выражения для $P(A)$ иллюстрируется на рис.2, где событие А условно изображено в виде заштрихованного круга, а гипотезы в виде секторов, имеющих пересечение с кругом.

$$P(A) = P(H_1A + H_2A + \ldots + H_nA) = P(H_1A) + P(H_2A) + \ldots + P(H_nA).$$

Применив к каждому из слагаемых правило умножения вероятностей, получим формулу полной вероятности

$$P(A) = P(H_1)P(A|H_1) + P(H_2)P(A|H_2) + \ldots + P(H_n)P(A|H_n). \qquad (1.12)$$

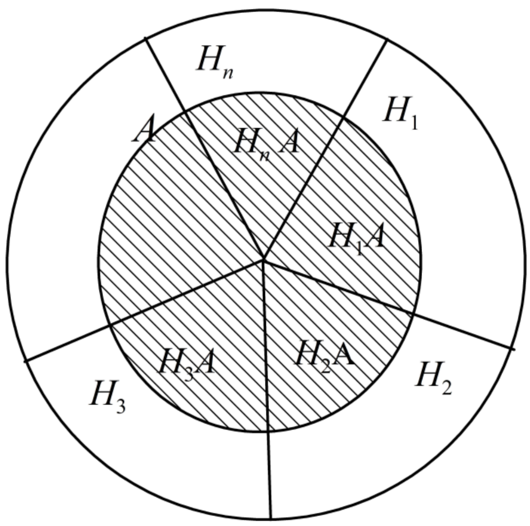


Рис. 1.2

Пример

Решено приобрести новый смартфон фирмы SmartX. Известно, что 60% смартфонов данной фирмы производят в Юго-Восточной Азии (включая континентальный Китай), 30% – в Европе, а 10% – это контрафактная продукция низкого качества. Известно также, что вероятность брака для смартфона, произведенного в ЮВА составляет 0,05, для смартфона произведенного в Европе 0,04, для контрафактной продукции – 0,5. Найти вероятность того, что приобретенный смартфон окажется бракованным.

Обозначим, как A событие, состоящее в том, что купленный смартфон оказался бракованным.

Априори можно сделать три предположения или гипотезы о приобретаемом изделии:

H_1 – смартфон изготовлен в ЮВА,

H_2 – смартфон изготовлен в Европе,

H_3 – это контрафактная продукция.

Вероятности гипотез известны: $P(H_1)=0{,}6$, $P(H_2)=0{,}3$, $P(H_3)=0{,}1$. Известны также условные вероятности $P(A\,|\,H_i)$. $P(A\,|\,H_i)$ – это вероятность брака при условии, что имела место гипотеза H_i: $P(A\,|\,H_1)=0{,}05$, $P(A\,|\,H_2)=0{,}04$, $P(A\,|\,H_3)=0{,}5$. Применяя формулу (1.12), получим

$$P(A)=0{,}6\cdot 0{,}05+0{,}3\cdot 0{,}04+0{,}1\cdot 0{,}5=0{,}092.$$

Очевидно, что данный результат можно интерпретировать и по-другому: 9,2% продающихся смартфонов данной фирмы являются бракованными.

1.3. Формула Байеса (формула гипотез)

Все обозначения предыдущего пункта остаются в силе. Но теперь предполагается, что опыт выполнен и как свершившийся факт имело место событие A. Требуется найти $P(H_j\,|\,A)$ – вероятность того, что опыт протекал в соответствии с гипотезой H_j при условии, что событие A произошло. Вероятности $P(H_j\,|\,A)$ называют апостериорными вероятностями, вероятности $P(H_j)$ – априорными.

Ответ задачи получается непосредственно из правила умножения вероятностей:

$$P(AB)=P(B)P(A\,|\,B)=P(A)P(B\,|\,A).$$

Просто заменим здесь B на H_j:

$$P(H_j)P(A\,|\,H_j)=P(A)P(H_j\,|\,A).$$

Отсюда получаем искомую апостериорную вероятность

$$P(H_j\,|\,A)=\frac{P(H_j)P(A\,|\,H_j)}{P(A)}=\frac{P(H_j)P(A\,|\,H_j)}{\sum\limits_{k=1}^{n}P(H_k)P(A\,|\,H_k)}. \tag{1.13}$$

Данная формула называется формулой Байеса (гипотез).

Пример

Вернемся к примеру предыдущего пункта. Смартфон куплен, он оказался не бракованным. Найти вероятность того, что он произведен в Европе.

Напомним, что как A мы обозначили событие, состоящее в том, что смартфон бракован. Таким образом, в результате опыта, который заключается в эксплуатации смартфона, имело место событие $\overline{A}$. Требуется найти вероятность $P(H_2 \mid \overline{A})$. Применив (1.13) получим

$$P(H_2 \mid \overline{A}) = \frac{P(H_2)P(\overline{A} \mid H_2)}{P(\overline{A})} = \frac{0{,}3 \cdot (1 - 0{,}04)}{1 - 0{,}092} \approx 0{,}317.$$

1.4. Формула Бернулли

Рассмотрим следующую задачу. Пусть производится n *независимых* опытов, в каждом из которых некоторое событие A появляется с одной и той же вероятностью p. Какова вероятность, что в n опытах событие A появится ровно m раз?

Обозначим искомую вероятность как P_m. На рисунке 3 схематично изображены m кружков – опытов в которых событие произошло и остальные $(n - m)$ опытов, в которых A не произошло, т.е. имело место событие $\overline{A}$ противоположное A.

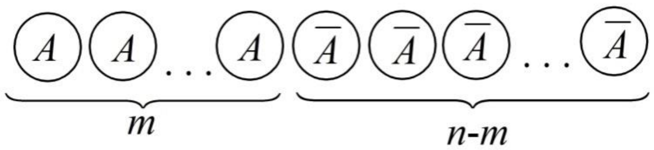


Рис. 1.3

Обозначим вероятность непоявления A как $q = 1 - p$. Применяя правило умножения вероятностей, с учетом того, что все опыты независимы, для вероятности реализации схемы, изображенной на рис.1.3, получим

$$\underbrace{p \cdot p \cdot \ldots \cdot p}_{m} \cdot \underbrace{q \cdot q \cdot \ldots \cdot q}_{n-m} = p^m q^{n-m}.$$

Однако схема рис.1.3 соответствует ситуации, когда событие A произошло в m первых опытах, в то время как по условию задачи данное событие может появляться в произвольных опытах. Число способов, которыми можно разместить m «кружков» A по n местам равно C_n^m, каждому такому сочетанию соответствует одна и та же вероятность $p^m q^{n-m}$. Применяя аксиому сложения вероятностей, окончательно получим

$$P_m = C_n^m p^m q^{n-m}, m = 0, 1, 2, \ldots, n. \quad (1.14)$$

Полученная формула называется формулой Бернулли или биноминальным распределением. Смысл термина «распределение» будет прояснен в следующей главе.

Сумма всех вероятностей P_m должна быть равна единице. Это действительно так, что легко показать, используя формулу, известную как бином Ньютона:

$$\sum_{m=0}^{n} C_n^m a^{n-m} b^m = (a + b)^n.$$

Теперь получим

$$\sum_{m=0}^{n} P_m = \sum_{m=0}^{n} C_n^m p^m q^{n-m} = (p+q)^n = 1.$$

Примеры.

1) Найти вероятность того, что при десяти подбрасываниях монеты герб появится пять раз.

Предполагается, что вероятность появления герба при подбрасывании монеты $p = 1/2$, поэтому искомая вероятность равна

$$P_5 = C_{10}^5 \frac{1}{2^5}\frac{1}{2^5} = \frac{6\cdot 7\cdot 8\cdot 9\cdot 10}{2\cdot 3\cdot 4\cdot 5}\cdot\frac{1}{2^{10}} = \frac{63}{256} \approx 0{,}246.$$

2) Игральный кубик подбрасывается шесть раз. Найти вероятность того, что шестерка появится

а) ровно два раза;

б) ровно один раз;

в) хотя бы один раз.

Полагаем, что вероятность появления шестерки при подбрасывании игрального кубика $p = 1/6$.

а) $P_2 = C_6^2 \dfrac{1}{6^2}\left(\dfrac{5}{6}\right)^4 = \dfrac{1}{2}\left(\dfrac{5}{6}\right)^5 \approx 0{,}201.$

б) $P_1 = C_6^1 \dfrac{1}{6}\left(\dfrac{5}{6}\right)^5 = \left(\dfrac{5}{6}\right)^5 \approx 0{,}402.$

в) Событие, состоящее в том, что шестерка появится хотя бы один раз ($m \ge 1$) противоположно событию, что шестерка не появится ни разу. Поэтому

$$P_{\ge 1} = 1 - P_0.$$

Заметим, что формула (1.14) справедлива и в случае, когда $m = 0$ (напомним здесь, что $0! = 1$). Но вероятность P_0 может быть найдена и по правилу умножения вероятностей как вероятность шести непоявлений шестерки.

$$P_{\ge 1} = 1 - \left(\frac{5}{6}\right)^6 \approx 0{,}665.$$

3) В партии деталей 20% бракованных. Для проверки отбирается пять деталей. Найти вероятность того, что среди отобранных ровно две бракованных. Рассмотреть два случая:

- общее число деталей равно 20,
- число деталей *очень* велико.

В первом случае число бракованных деталей равно четырем. Искомую вероятность можно найти, применив формулу (1.7)

$$P_2 = \frac{C_4^2 C_{16}^3}{C_{20}^5} = \frac{70}{323} \approx 0{,}217.$$

Во втором случае можно применить формулу Бернулли, $p = 0{,}2 = 1/5$:

$$P_2 = C_5^2 \frac{1}{5^2}\left(\frac{4}{5}\right)^3 = \frac{128}{625} \approx 0{,}205.$$

В первом случае формула Бернулли не применима, так как число деталей конечно и невелико. Условие, что в каждом опыте событие A (деталь бракована) появляется с одной и той же вероятностью p, не выполняется. Вероятность того, что деталь бракована, равна 1/5 только для первой детали, взятой для проверки. Для второй детали эта вероятность будет равна 4/19 или 3/19 в зависимости от того небракованной или бракованной оказалась первая деталь.

Во втором случае, так как число деталей очень велико, вероятность p можно с хорошей точностью считать постоянной. Поэтому, хотя и не вполне точно, схема Бернулли применима.

Задачи для самостоятельного решения

1) Из пяти карточек разрезной азбуки составлено слово «карта». Ребенок, не умеющий читать, рассыпал карточки, затем сложил их в случайном порядке. Найти вероятность того, что у него снова получится слово «карта».

2) Среди 100 лотерейных билетов 5 выигрышных. Найти вероятность того, что

- 2 наудачу выбранных билета окажутся выигрышными;
- из пяти купленных билетов только один окажется выигрышным.

3) Вероятность хотя бы одного попадания в цель при трех выстрелах равна 0,973. Найти вероятность попадания в цель при одном выстреле.

4) Чеканщик кладет в каждый ящик вместимостью n монет одну фальшивую. Для проверки берется партия из n ящиков и из каждого ящика проверяют одну монету. Полагая, что n очень велико найти приближенно вероятность того, что чеканщик не будет разоблачен (все проверенные монеты окажутся не фальшивыми).

5) На поверхность, разграфленную на квадраты, бросается монета. Сторона квадрата 4 см, радиус монеты 1,5 см. Найти вероятность того, что монета окажется *полностью* внутри одного из квадратов.

6) В первой урне 6 белых шаров и 4 черных. Во второй урне 5 белых и 7 черных. Из первой урны во вторую некто переложил один шар неизвестно какого цвета. Какова вероятность вытащить после этого белый шар из второй урны. В задачах по теории вероятностей под урной подразумевается черный ящик, содержимое которого, возможно, известно, но не наблюдаемо.

7) В группе из 30 студентов, пришедших на экзамен, 8 студентов подготовлены отлично, 10 – хорошо, 8 – удовлетворительно, 4 – неудовлетворительно. Общее количество экзаменационных вопросов – 20. Отлично подготовленный студент знает ответы на все 20 вопросов, хорошо подготовленный – на 16, удовлетворительно подготовленный – на 10, неудовлетворительно подготовленный – на 5. Вызванный студент ответил на все три вопроса экзаменационного билета. Какова вероятность, что этот студент подготовлен

- отлично;
- неудовлетворительно.

8) При игре в шахматы с равносильным противником вероятность выигрыша равна $p_1 = 1/4$, вероятность проигрыша $p_2 = 1/4$, вероятность ничьей $p_3 = 1/2$. Будет сыграно восемь партий. Найти вероятности следующих событий

- выиграть ровно две партии;
- выиграть половину партий;
- половина партий будет сыграна вничью.

9) Производится стрельба по летящей мишени. Вероятность попадания при одном выстреле равно 0,6. При одном попадании вероятность поражения мишени равна 0,5, при двух попаданиях 0,9, при более чем двух попаданиях – единица. Производится пять выстрелов. Найти вероятность того, что мишень будет поражена.

Ответы.

1)1/60 2) 1/495, 0,211 3) 0,7 4) $1/e \approx 0{,}368$ 5) 1/16 6) 28/65 7) 0,59; 0,003 8)0,311; 0,087; 0,273 9) 0,928.

Глава 2. Случайные величины

2.1. Законы распределения

Прежде всего, дадим определение случайной величины.

Случайной величиной называется величина, которая в опыте со случайным исходом принимает те или иные точно не предсказуемые значения.

Множество всех возможных значений случайной величины далее будем обозначать буквой T.

Примеры.

1) Бросается игральная кость. Случайная величина – число выпавших очков. $T = \{1, 2, 3, 4, 5, 6\}$.

2) Бросаются три монеты. Случайная величина – число выпавших гербов. $T = \{0, 1, 2, 3\}$.

3) С конвейера сходят некоторые изделия и проверяются на наличие брака. Случайная величина – количество проверенных изделий до появления первого бракованного. Множество возможных значений – множество неотрицательных целых чисел: $T = \{0, 1, 2, \ldots\}$.

4) Между пунктами А и Б протянута телеграфная линия длина которой равна L. Случайная величина – расстояние от пункта А до точки возможного обрыва. T – множество действительных чисел от нуля до L: $T = \{x : 0 \le x \le L\}$.

5) Напряжение U в сети переменного тока: $T = \{u : 0 \le u < \infty\}^{*}$.

Из приведенных примеров видим, что случайные величины бывают двух принципиально разных видов.

Дискретными случайными величинами называют величины множество возможных значений которых конечно (примеры 1, 2) или бесконечно (пример 3), но *счетно*.

Непрерывными случайными величинами называют величины, множество возможных значений которых сплошь занимает некоторый промежуток числовой прямой и, следовательно, *несчетно* (примеры 4, 5).

Далее случайными величины будем обозначать прописными буквами, их определенные значения – строчными.

Законом распределения случайной величины или просто распределением называется любое правило, по которому каждому возможному значению величины сопоставляется вероятность появления этого значения.

Теперь становится понятным смысл названия «биноминальное распределение». В схеме Бернулли рассматривается случайная величина – число появлений некоторого события A в независимых испытаниях. Формула (1.14) сопоставляет каждому возможному значению m (m=0,1,2,...,n) этой случайной величины – вероятность P_m появления этого значения.

Для дискретных случайных величин закон распределения часто задают в виде ряда распределения.

---

[*] Здесь некоторое недоумение может вызвать верхняя граница множества, но в контексте рассматриваемых здесь примеров это не принципиально. Можно считать, например, что напряжение не может превысить 1000 В.

Ряд распределения это таблица в первой строке которой перечислены все возможные значения случайной величины в порядке возрастания, во второй – соответствующие вероятности.

Для примера 1 (число выпавших очков при бросании игральной кости) ряд распределения имеет вид

Таблица 2.1

X	1	2	3	4	5	6
P	1/6	1/6	1/6	1/6	1/6	1/6

Вероятности в примере 2 легко найти по формуле (1.14) для n=3, p=1/2. Таким образом, ряд распределения в данном примере будет иметь вид

Таблица 2.2

X	0	1	2	3
P	1/8	3/8	3/8	1/8

Важно иметь в виду неотъемлемое свойство ряда распределения: все вероятности во второй строке таблицы должны в сумме давать единицу. Это очевидно, так как в первой строке перечислены все возможные значения случайной величины.

Для непрерывных случайных величин ряд распределения нельзя составить в принципе. В этом случае используют функцию распределения.

Функцией распределения $F(x)$ случайной величины X называется вероятность того, что X примет значение строго меньшее заданного значения x:

$$F(x) = P(X < x). \tag{2.1}$$

Заметим, что функцию распределения можно записать и для дискретной случайной величины, но обычно этого не делают, так как ряд распределения в таком случае удобнее.

Из самого определения функции распределения вытекают следующие ее свойства:

1) $\lim\limits_{x\to-\infty} F(x) = 0$,

2) $\lim\limits_{x\to+\infty} F(x) = 1$,

3) $F(x)$ является неубывающей функцией: если $x_2 > x_1$, $F(x_2) \geq F(x_1)$.

В качестве иллюстративного примера составим функцию распределения для примера 1.

При $x \leq 1$ функцию распределения следует положить равной нулю, $F(x) = 0$, так как выпадение числа очков меньше единицы является, очевидно, невозможным событием.

Рассмотрим промежуток $(1, 2]$. Возьмем любое число x из этого промежутка. Условию $X < x$ удовлетворяет единственное событие $X = 1$, вероятность появления которого равна 1/6. Таким образом, при $x \in (1, 2]$ $F(x) = 1/6$.

Пусть далее $x \in (2, 3]$. Теперь условию $X < x$ удовлетворяют два *несовместных* события $X = 1$, $X = 2$. Таким образом, в данном промежутке $F(x) = \dfrac{1}{6} + \dfrac{1}{6} = \dfrac{1}{3}$.

Рассуждая подобным образом, получим

$x \in (3, 4]$, $F(x) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}$,

$x \in (4, 5]$, $F(x) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{2}{3}$,

$x \in (5, 6]$, $F(x) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{5}{6}$,

$x \in (6, +\infty)$, $F(x) = 1$.

График функции изображен на рис. 2.1. Видим, что основные свойства функции распределения имеют место. Непрерывность не является неотъемлемым свойством функции распределения. Она может иметь счетное множество разрывов первого рода.

Для дискретной случайной величины величина скачка функции $F(x)$ в некоторой точке $x = a$ равна вероятности, что случайная величина примет значение a.

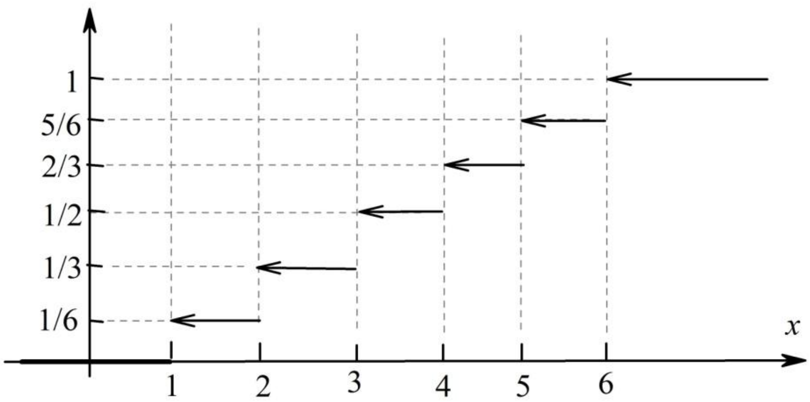


Рис. 2.1. График функции $F(x)$ для случайной величины – число выпавших очков при бросании игрального кубика.

Для непрерывной случайной величины функция распределения является непрерывной функцией.

Решим теперь следующую задачу. Пусть для случайной величины X известна функция распределения $F(x)$. Нужно найти вероятность, события $a \le X < b$, т.е. величину $P(a \le X < b)$. Рассмотрим событие $\{X < b\}$, состоящее в том, что X примет значение строго меньшее, чем b, иначе множество точек, лежащих слева от точки b на числовой прямой. Это событие (множество) можно представить как сумму двух несовместных событий (непересекающихся множеств): $\{X < b\} = \{X < a\} + \{a \le X < b\}$. Используя аксиому сложения, получим

$$P(X < b) = P(X < a) + P(a \le X < b),$$

$$P(a \le X < b) = P(X < b) - P(X < a).$$

Воспользовавшись определением (2.1) приходим к выражению

$$P(a \le X < b) = F(b) - F(a). \tag{2.2}$$

Найдем вероятность того, что X примет строго определенное значение: $X = a$. Данную вероятность попытаемся найти как предел

$$P(X=a)=\lim_{\Delta x\to 0}P(a\le X<a+\Delta x)=\lim_{\Delta x\to 0}F(a+\Delta x)-F(a).$$

Для дискретной случайной величины этот предел равен скачку, который функция $F(x)$ испытывает в точке a, то есть, собственно вероятности $P(X=a)$. Результат тривиален.

Но для непрерывной случайной величины, функция распределения которой непрерывна, имеем

$$P(X=a)=\lim_{\Delta x\to 0}F(a+\Delta x)-F(a)=F(a)-F(a)=0.$$

Пришли к парадоксальному на первый взгляд результату. Вероятность того, что непрерывная случайная величина примет строго определенное значение равна нулю. На самом деле парадокса здесь нет. Пусть, например, мы измеряем напряжение в бытовой сети переменного тока. Какова вероятность, что значение напряжения ровно 220 В? Пусть используемый нами вольтметр с ценой деления 1 В показал значение ровно 220 В. Но значит ли это, что напряжение действительно равно этому значению? Разумеется, нет. Более точный вольтметр даст значение, допустим, 219,6 В. Взяв еще более точный прибор, получим новое, более точное значение и т.д.

Таким образом, полученный результат просто отражает саму природу непрерывной случайной величины, множество возможных значений которой *несчетно*. Образно говоря, максимальное значение вероятности – единица мы распределили по несчетному множеству точек, получив, таким образом, ноль в каждой из них.

Но с другой стороны на интуитивном уровне мы понимаем, что вероятность получить напряжение 220 В много выше, чем, например, напряжение 400 В, а наши рассуждения дают вероятность ноль для обоих значений. Дело в том, что для непрерывной случайной величины корректный вопрос о значении вероятности должен быть поставлен следующим образом: какова вероятность, что непрерывная случайная величина примет значение из некоторого промежутка? Вероятность того, что измеряемое напряжение примет значение в промежутке от 220 до 221 В много выше, чем вероятность попадания его в промежуток от 400 до 401 В.

В связи со сказанным следует отметить следующее обстоятельство. При вычислении вероятности попадания *непрерывной* случайной величины в некоторый промежуток абсолютно неважно включаются ли в рассмотрение границы промежутка, т.е. используется ли знаки «строго меньше» или «меньше или равно», «строго больше» или «больше или равно».

Пусть имеется непрерывная случайная величина X с функцией распределения $F(x)$.

Плотностью распределения (вероятности) называется функция

$$f(x)=F'(x). \qquad (2.3)$$

Таким образом, используя формулу (2.2), можно записать

$$f(x)=\lim_{\Delta x\to 0}\frac{F(x+\Delta x)-F(x)}{\Delta x}=\lim_{\Delta x\to 0}\frac{P(x\le X<x+\Delta x)}{\Delta x}.$$

Для малых Δx $f(x)\approx\dfrac{P(x\le X<x+\Delta x)}{\Delta x}$. Обычная плотность – это масса, отнесенная к объему, который тело данной массы занимает. Плотность тела может меняться от точки к точки. Чем меньше величина выбранного объема, тем с большей точностью мы найдем плотность в данной точке. Плотность распределения вероятности – это вероятность попадания случайной величина в заданный интервал, отнесенная к величине интервала Δx,

что объясняет использование термина «плотность» в данном случае. Чем меньше Δx, тем точнее мы найдем $f(x)$ в данной точке.

Рассмотрим интеграл по промежутку $[a,b]$ от плотности распределения. Так как из самого определения (2.3) следует, что функция распределения является первообразной для плотности распределения, используя формулу Ньютона-Лейбница и (2.2), получим

$$\int_a^b f(x)dx = F(b) - F(a) = P(a \le X < b). \tag{2.4}$$

Отсюда следует, что площадь под графиком функции $f(x)$ на данном промежутке равна вероятности попадания случайной величины в этот промежуток – см. рис.2.1а.

Рассмотрим теперь несобственный интеграл

$$\int_{-\infty}^x f(t)dt = F(x) - \lim_{a \to -\infty} F(a),$$

так как в силу свойства функции $F(x)$ предел в правой части равенства равен нулю, имеем

$$\int_{-\infty}^x f(t)dt = F(x). \tag{2.5}$$

Несложно получить свойство

$$\int_{-\infty}^{\infty} f(x)dx = 1, \tag{2.6}$$

которое аналогично свойству «сумма всех вероятностей равна единице» для ряда распределения.

Из (2.6) следует, что площадь под графиком плотности распределения равна единице – см. рис. 2.1б.

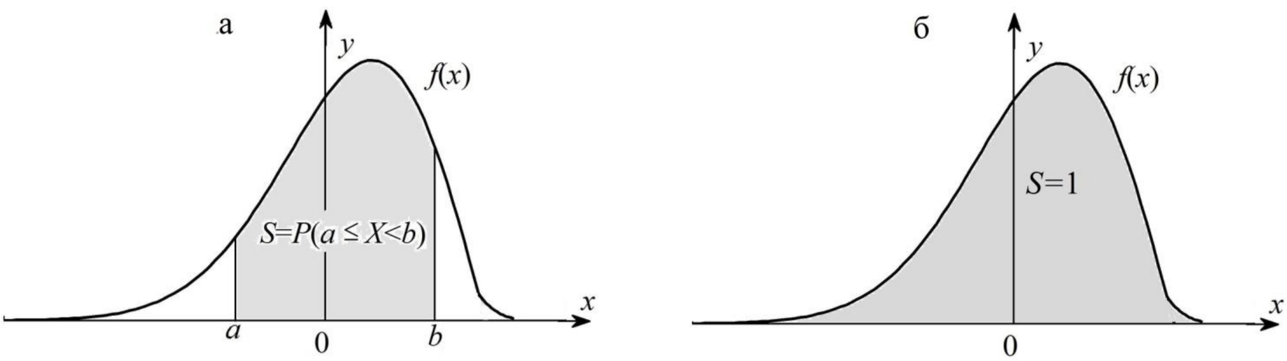


Рис.2.1

Пример.

Найдем функцию распределения для примера 4 из данного пункта. В предположении, что вероятность обрыва на некотором участке длины Δx не зависит от расположения этого участка, функция распределения имеет вид

$$F(x) = \begin{cases} 0, & x \in (-\infty, 0] \\ \dfrac{x}{L}, & x \in (0, L] \\ 1, & x \in (L, \infty) \end{cases}.$$

Плотность распределения легко найти дифференцированием

$$f(x)=\begin{cases}0, & x\in(-\infty,0]\\ \dfrac{1}{L}, & x\in(0,L]\\ 0, & x\in(L,\infty)\end{cases}.$$

Плотность распределения имеет вид ступеньки или прямоугольника, площадь которого равна единице.

2.2 Числовые характеристики случайных величин

Пусть имеется дискретная случайная величина *X*, заданная рядом распределения

X	x_1	x_2	....	x_i	...
P	p_1	p_2	....	p_i	...

Математическим ожиданием дискретной случайной величины *X* называется сумма произведений всех значений *X* на соответствующие вероятности

$$M[X]=\sum_i x_i p_i. \qquad (2.7)$$

Число слагаемых в сумме может быть как конечным так и бесконечным в зависимости от того конечным или бесконечным является число возможных значений данной случайной величины.

Если в формуле (2.7) заменить вероятности на частоты появления значений случайной величины в n опытах или измерениях, то математическое ожидание станет средним арифметическим наблюденных значений случайной величины. Данный факт позволяет интерпретировать математическое ожидание как *среднее значение* случайной величины. Но не следует отождествлять понятия среднее значение и среднее арифметическое. Фактически, среднее значение это просто синоним термина математическое ожидание, которое определяется формулой (2.7) и приведенной ниже формулой (2.8) для непрерывной случайной величины. Но среднее арифметическое и математическое ожидание тесно связаны друг с другом. Характер этой связи будет прояснен в главе 3.

Для непрерывной случайной величины с плотностью распределения $f(x)$ математическое ожидание определяется как

$$M[X]=\int_{-\infty}^{\infty} x f(x)dx. \qquad (2.8)$$

В формулах вместо довольно громоздкого обозначения $M[X]$ бывает удобнее пользоваться обозначением m_x: $m_x \equiv M[X]$.

Величиной, показывающей, как сильно случайная величина может отклоняться от своего математического ожидания (т.е. насколько сильно рассеяна случайная вокруг своего среднего значения), является *дисперсия*.

Дисперсию можно определить как математическое ожидание квадрата отклонения случайной величины от своего среднего значения:

$$D[X]=M\left[(X-m_x)^2\right]. \qquad (2.9)$$

Отсюда для дискретной случайной величины

$$D[X]=\sum_i (x_i-m_x)^2 p_i=\sum_i x_i^2 p_i-2\sum_i m_x x_i p_i+\sum_i m_x^2 p_i=\sum_i x_i^2 p_i-2m_x\sum_i x_i p_i+m_x^2\sum_i p_i=$$
$$=\sum_i x_i^2 p_i-2m_x^2+m_x^2=\sum_i x_i^2 p_i-m_x^2=M\left[X^2\right]-M^2[X],$$

где мы учли, что сумма всех вероятностей равна единице: $\sum_i p_i=1$.

Для непрерывной случайной величины

$$D[X]=\int_{-\infty}^{\infty}(x-m_x)^2 f(x)dx=\int_{-\infty}^{\infty}x^2 f(x)dx-2m_x\int_{-\infty}^{\infty}x^2 f(x)dx+m_x^2\int_{-\infty}^{\infty}f(x)dx=$$
$$=\int_{-\infty}^{\infty}x^2 f(x)dx-m_x^2=M\left[X^2\right]-M^2[X],$$

где учтено, что $\int_{-\infty}^{\infty}f(x)dx=1$.

Таким образом, дисперсия может быть определена как

$$D[X]=M\left[(X-m_x)^2\right]=M\left[X^2\right]-M^2[X]. \qquad (2.10)$$

Оба выражения здесь совершенно равнозначны, но при практических вычислениях, как правило, удобнее второе.

Альтернативным обозначением для дисперсии $D[X]$ является D_x.

Дисперсия имеет размерность квадрата размерности самой случайной величины. Поэтому использовать ее как характеристику отклонения случайной величины от среднего значения не совсем удобно. Вместо дисперсии часто пользуются средним квадратическим отклонением, которое определяется просто как квадратный корень из дисперсии. Для случайной величины X среднее квадратическое отклонение обозначается как σ_x:

$$\sigma_x=\sqrt{D[X]}. \qquad (2.11)$$

Заметим здесь, что дисперсия как математическое ожидание неотрицательной величины не может быть отрицательной.

Легко доказать следующее свойство математического ожидания и дисперсии

$$\mathrm{M}[cX]=c\mathrm{M}[X],\ \mathrm{D}[cX]=c^2\mathrm{D}[X],\ \text{где}\ c=const.$$

Примеры.

1) По многолетним данным в ВУЗе получено следующее распределение оценок (пятибалльная система) по дисциплине математика

Оценка X	2	3	4	5
Вероятность P	0,05	0,40	0,35	0,20

Найти математическое ожидание, дисперсию и среднее квадратическое отклонение оценки X.

Математическое ожидание в соответствии с формулой (2.7) равно

$$M[X]=2\cdot 0{,}05+3\cdot 0{,}40+4\cdot 0{,}35+5\cdot 0{,}20=3{,}7.$$

Для вычисления дисперсии используем вторую из формул (2.10)

$D[X] = 2^2 \cdot 0{,}05 + 3^2 \cdot 0{,}40 + 4^2 \cdot 0{,}35 + 5^2 \cdot 0{,}20 - 3{,}7^2 = 0{,}71.$

Среднее квадратическое отклонение:

$\sigma_x = \sqrt{0{,}71} \approx 0{,}84$.

2) С конвейера сходят изделия. Рассматривается случайная величина X – число проверенных изделий до появления первой бракованной. Найти математическое ожидание и дисперсию данной случайной величины. Вероятность брака известна и равна p.

Обозначим $q = 1 - p$. Прежде всего, составим ряд распределения случайной величины X:

X	0	1	2	....	n	...
P	p	pq	pq^2	....	pq^n	...

Вероятности во второй строке таблицы образуют бесконечную геометрическую прогрессию первый член и знаменатель которой равны p и q соответственно, причем по смыслу задачи следует считать, что $0 < q < 1$. Поэтому для суммы вероятностей получим

$$p + pq + pq^2 + \ldots + pq^n + \ldots = \frac{p}{1-q} = 1,$$

что является необходимым условием корректности ряда распределения.

Для математического ожидания имеем

$$M[X] = \sum_{n=0}^{\infty} npq^n = pq\sum_{n=1}^{\infty} nq^{n-1} = pq\sum_{n=1}^{\infty}\frac{d}{dq}q^n = pq\frac{d}{dq}\sum_{n=1}^{\infty}q^n = pq\frac{d}{dq}\frac{q}{1-q} = pq\frac{1}{(1-q)^2} =$$

$$= \frac{q}{p} = \frac{1-p}{p}.$$

Здесь мы использовали тот факт, что равномерно сходящиеся ряды можно дифференцировать почленно.

Находим дисперсию

$$D[X] = \sum_{n=0}^{\infty} n^2 pq^n - \left(\frac{q}{p}\right)^2 = \sum_{n=1}^{\infty} n(n-1)pq^n + \sum_{n=1}^{\infty} npq^n - \left(\frac{q}{p}\right)^2 = pq^2\sum_{n=1}^{\infty}\frac{d^2}{dq^2}q^n + \frac{q}{p} - \left(\frac{q}{p}\right)^2 =$$

$$= pq^2\frac{d^2}{dq^2}\sum_{n=1}^{\infty}q^n + \frac{q}{p} - \left(\frac{q}{p}\right)^2 = pq^2\frac{d^2}{dq^2}\frac{q}{1-q} = pq^2 \cdot \frac{2}{(1-q)^3} + \frac{q}{p} - \left(\frac{q}{p}\right)^2 = \frac{2q^2}{p^2} + \frac{q}{p} - \left(\frac{q}{p}\right)^2 =$$

$$= \frac{q}{p} + \frac{q^2}{p^2} = \frac{1-p}{p^2}.$$

3) На рисунке 2.2 показана плотность распределения некоторой случайной величины X. Найти функцию распределения, математическое ожидание, дисперсию и среднее квадратическое отклонение.

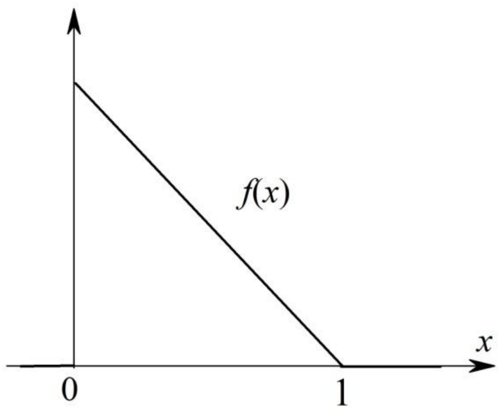


Рис. 2.2.

Так как площадь под графиком плотности распределения должна быть равна единице, то $f(0)=2$. Теперь легко записать аналитическое выражение для $f(x)$:

$$f(x)=\begin{cases}0,\ x<0\\ 2-2x,\ 0\le x\le 1\\ 0,\ x>1\end{cases}.$$

В промежутке от нуля до единицы функцию распределения в соответствии с (2.5) легко найти интегрированием:

$$F(x)=\int_0^x (2-2t)dt=2x-x^2.$$

Таким образом

$$F(x)=\begin{cases}0,\ x<0\\ 2x-x^2,\ 0\le x\le 1.\\ 1,\ x>1\end{cases}$$

Так как плотность распределение равна нулю вне промежутка $[0,1]$ несобственный интеграл (2.8) становится интегралом по конечному промежутку. Тогда

$$m_x=\int_{-\infty}^{\infty} xf(x)dx=\int_0^1 x(2-2x)dx=\left(x^2-\frac{2}{3}x^3\right)\Bigg|_0^1=\frac{1}{3},$$

$$D_x=\int_{-\infty}^{\infty} x^2 f(x)dx-m_x^2=\int_0^1 x^2(2-2x)dx-\frac{1}{9}=\left(\frac{2}{3}x^3-\frac{1}{2}x^4\right)\Bigg|_0^1-\frac{1}{9}=\frac{1}{18},$$

$$\sigma_x=\frac{1}{3\sqrt{2}}.$$

Среднее значение и дисперсия случайной величины не являются единственными числовыми характеристиками случайной величины. В теории вероятностей существует понятие «момент порядка k», который для случайной величины X определяется как

$$\mu_k[X]=M\left[X^k\right]. \tag{2.12}$$

В соответствии с формулами (2.7), (2.8) для дискретной случайной величины

$$\mu_k[X]=\sum_i x_i^k p_i, \tag{2.13a}$$

для непрерывной

$$\mu_k[X] = \int_{-\infty}^{\infty} x^k f(x)dx. \tag{2.13б}$$

Центральным моментом порядка k называется величина

$$\bar{\mu}_k[X] = M\left[(X - m_x)^k\right]. \tag{2.14}$$

Таким образом, дисперсия является центральным моментом второго порядка.

2.3. Некоторые законы распределения

Здесь мы рассмотрим законы распределения, которые часто возникают в задачах теории вероятностей.

1. *Биноминальное распределение*. Задача, в которой возникает данное распределение, уже рассматривалась в п.1.4. Само распределение определяется формулой (1.14). Найдем математическое ожидание и дисперсию случайной величины X – числа появления некоторого события в n независимых опытах.

$$M[X] = \sum_{k=0}^{n} kC_n^k p^k q^{n-k} = \sum_{k=1}^{n} \frac{k\,n!}{k!(n-k)!} p^k q^{n-k} n = \sum_{k=1}^{n} \frac{n!}{(k-1)!((n-1)-(k-1))!} p^k q^{n-k} =$$

$$= np\sum_{k=1}^{n} \frac{(n-1)!}{(k-1)!((n-1)-(k-1))!} p^{k-1} q^{(n-1)-(k-1)}.$$

Обозначим $l = k-1$. Используя бином Ньютона, получим

$$M[X] = np\sum_{l=0}^{n-1} C_{n-1}^l p^l q^{(n-1)-l} = np(p+q)^{n-1},$$

$$M[X] = np. \tag{2.15}$$

$$D[X] = \sum_{k=0}^{n} k^2 C_n^k p^k q^{n-k} - n^2 p^2.$$

$$\sum_{k=0}^{n} k^2 C_n^k p^k q^{n-k} = \sum_{k=1}^{n} \frac{kn!}{(k-1)!(n-k)!} p^k q^{n-k} =$$

$$= \sum_{k=1}^{n} \frac{(k-1)n!}{(k-1)!(n-k)!} p^k q^{n-k} + p\sum_{k=1}^{n} \frac{n!}{(k-1)!(n-k)!} p^{k-1} q^{n-k} = \sum_{k=2}^{n} \frac{(k-1)n!}{(k-1)!(n-k)!} p^k q^{n-k} + np =$$

$$= n(n-1)p^2 \sum_{k=2}^{n} \frac{(n-2)!}{(k-2)!(n-2-(k-2))!} p^{k-2} q^{n-k} + np = n(n-1)p^2 \sum_{l=0}^{n-2} \frac{(n-2)!}{l!(n-2-)!} p^l q^{n-2-l} + np =$$

$$= n(n-1)p^2 + np.$$

$$D[X] = n(n-1)p^2 + np - n^2p^2 = np - np^2 = np(1-p),$$

$$D[X] = npq. \tag{2.16}$$

2. *Распределение Пуассона*. Предположим, что число испытаний n в схеме Бернулли настолько велико, что практически его можно считать бесконечным, а вероятность p настолько мала, что практически ее можно считать равной нулю. Выполним предельный переход $n \to \infty$, $p \to 0$ в выражении (1.14), но потребовав, чтобы произведение np оставалось постоянной величиной:

$$np = a = const.$$

Подставив в (1.14) $p = a/n$ найдем

$$\lim_{n\to\infty} C_n^m \left(\frac{a}{n}\right)^m \left(1-\frac{a}{n}\right)^{n-m} = \lim_{n\to\infty} \frac{n(n-1)\cdot\ldots\cdot(n-m+1)}{n^m}\frac{a^m}{m!}\frac{(1-a/n)^n}{(1-a/n)^m} =$$

$$= \frac{a^m}{m!}\lim_{n\to\infty}\left(1\cdot\left(1-\frac{1}{n}\right)\cdot\left(1-\frac{2}{n}\right)\cdot\ldots\cdot\left(1-\frac{m-1}{n}\right)\right)\cdot\lim_{n\to\infty}\frac{1}{(1-a/n)^m}\cdot\lim_{n\to\infty}\left[(1-a/n)^{n/a}\right]^a.$$

Первый и второй пределы в данном выражении равны единице, а третий предел несложно найти. Он равен e^{-a}. Итак, окончательно получим следующую формулу для распределения Пуассона

$$P_m = \frac{a^m}{m!}e^{-a}, \; m = 0,1,2,\ldots \tag{2.17}$$

Так как для распределения Бернулли математическое ожидание равно произведению *np*, для распределения Пуассона оно равно параметру *a*:

$$M[X] = a. \tag{2.18}$$

Для дисперсии имеем

$$D[X] = \lim_{n\to\infty} npq = \lim_{n\to\infty} n\frac{a}{n}\left(1-\frac{a}{n}\right) = a\lim_{n\to\infty}\left(1-\frac{a}{n}\right) = a. \tag{2.19}$$

Таким образом, для распределения Пуассона математическое ожидание и дисперсия равны.

Используя разложение в ряд показательной функции e^x несложно показать, что сумма всех вероятностей (2.17) равна единице.

Пример.

В каждом опыте событие *A* появляется с вероятностью $p = 1/10$. Найти вероятность того, что в десяти независимых опытах событие *A* появится два раза.

Точный ответ дает здесь, конечно, биноминальное распределение:

$$P_2 = C_{10}^2 \frac{1}{10^2}\left(\frac{9}{10}\right)^8 = \frac{1}{2}\left(\frac{9}{10}\right)^9 \approx 0{,}194.$$

Теперь применим распределение Пуассона, которое выполняется здесь только приближенно, так как *n* не очень велико, а *p* не очень мало. Параметр *a* следует взять равным $a = 10\cdot 0{,}1 = 1$.

$$P_2 = \frac{1^2}{2!}e^{-1} = \frac{1}{2e} \approx 0{,}184$$

Видим, что точное и примерное значения вероятности хотя и не очень близки, но радикальной разницы между ними нет.

Понятие о пуассоновских (простейших) потоках.

Потоком называется последовательность одинаковых событий, происходящих в случайные, вообще говоря, моменты времени.

Важнейшей характеристикой потока является интенсивность λ, которая определяется как *среднее* число событий за единицу времени. Единицей измерения интенсивности является «время$^{-1}$».

Поток называется пуассоновским или простейшим, если выполняются следующие условия.

1) *Ординарность*. Это свойство заключается в том, что вероятность попадания на достаточно малый временной интервал Δt двух или более событий является пренебрежимо малой величиной по сравнению с вероятностью попадания на данный интервал одного события. Таким образом, ординарность означает, что события возникают поодиночке, а не группами по два, по три и т.д.

2) *Отсутствие последействия*. Возьмем два произвольных, но не перекрывающихся временных интервала, не обязательно равных и не обязательно смежных, τ_1 и τ_2 (см.рис.2.3). Отсутствие последействия означает, что число событий, произошедших за интервал τ_1, никак не повлияет на число событий, произошедших за τ_2. Фактически, это означает, что события не влияют друг на друга.

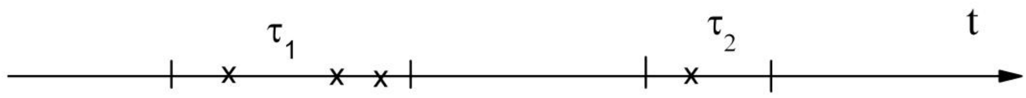


Рис. 2.3

3) *Стационарность*. Данное свойство означает, что вероятностные характеристики потока не зависят от времени. Подчеркнем, что речь идет именно о вероятностных характеристиках. Например, число событий, появляющихся за определенный промежуток времени, не является величиной постоянной в силу случайного характера потока. Но такая характеристика потока как *среднее* число событий за единичный интервал времени т.е. интенсивность остается величиной постоянной для стационарных потоков.

Смысл термина «пуассоновский поток» заключается в следующем. Разобьем некоторый временной промежуток t на очень большое, практически бесконечное, число малых интервалов $\Delta t = t / n$. В силу свойства 1) на малом интервале Δt может произойти не более, чем одно событие. Следовательно, каждый малый временной интервал фактически представляет собой испытание, в результате которого появляется или не появляется некоторое событие. В силу свойства 2) все испытания являются независимыми. Поэтому, учитывая, что общее число испытаний практически бесконечно, вероятность появления на данном временном промежутке t определенного числа событий будет определяться распределением Пуассона, откуда собственно и произошло название потока. В силу свойства 3) интенсивность λ не зависит от времени. *Среднее* число событий за время t равно λt, следовательно, параметр a в распределении Пуассона следует положить равным λt (a является математическим ожиданием для распределения Пуассона). Таким образом, вероятность того, что за время t произойдет ровно m событий, равна

$$P_m = \frac{(\lambda t)^m}{m!} e^{-\lambda t}. \tag{2.20}$$

Следует сказать, что пуассоновский поток является идеализацией, редко выполняющейся в чистом виде для реальных потоков.

Пример.

На автомобильную заправочную станцию (АЗС) поступает поток автомобилей с интенсивностью λ=45 час$^{-1}$. Полагая, что поток автомобилей является пуассоновским, найти вероятность того, что за две минуты на АЗС

а) не прибудет ни одного автомобиля,

б) прибудет два автомобиля,

в) прибудет хотя бы один автомобиль.

Произведение λt равно $\lambda t = 45 \cdot \frac{1}{60} \cdot 2 = 1{,}5$.

Поэтому

а) $P_0 = \frac{1{,}5^0}{0!} e^{-1{,}5} = \frac{1}{e^{1{,}5}} \approx 0{,}223$,

б) $P_2 = \frac{1{,}5^2}{2!} e^{-1{,}5} \approx 0{,}251$,

в) $P_{\geq 1} = 1 - P_0 \approx 0{,}777$.

3. *Показательное (экспоненциальное) распределение*.

Показательное распределение тесно связано с распределением Пуассона. Рассмотрим непрерывную случайную величину T – промежуток времени между двумя случайными событиями потока. Найдем функцию распределения $F(t)$ данной случайной величины. Напомним, что $F(t)$ по определению представляет собой вероятность того, что T примет значение строго меньшее чем t : $F(t) = P(T < t)$. Для того, чтобы выполнялось неравенство $T<t$ необходимо чтобы хотя бы одно событие потока попало на временной интервал t. Эта ситуация символически изображена на рис. 2.4

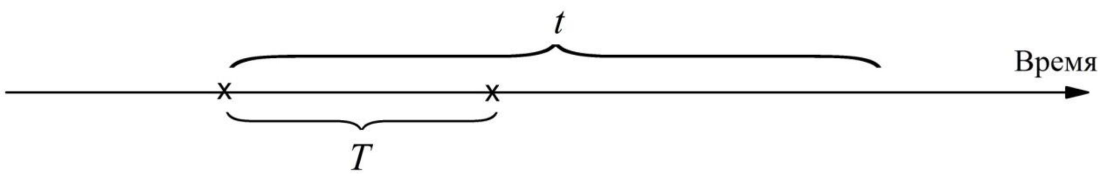


Рис.2.4

Событие, заключающееся в том, что на участок длины t попадет хотя бы одно событие, противоположно событию, состоящему в том, что на данный участок не попадет ни одного события. Следовательно

$$P(T < t) = 1 - P_0,$$

где P_0 вероятность того, что на промежутке t не произойдет ни одного события. Пусть поток событий является пуассоновским. Тогда, в соответствии с (2.20):

$$P_0 = e^{-\lambda t} .$$

Полагая, что переменная t, которая по смыслу является промежутком времени между двумя событиями, является величиной неотрицательной, приходим к распределению, называемому показательным или экспоненциальным

$$F(t)=\begin{cases}0, & t\le 0\\ 1-e^{-\lambda t}, & t>0\end{cases}. \qquad (2.21)$$

Для плотности распределения $f(t)=F'(t)$ получим выражение

$$f(t)=\begin{cases}0, & t\le 0\\ \lambda e^{-\lambda t}, & t>0\end{cases}. \qquad (2.22)$$

Несложно найти математическое ожидание случайной величины *T*, распределенной по показательному закону, то есть среднее значение промежутка времени между двумя случайными событиями:

$$M[T]=\int_0^\infty t\,f(t)dt=\lambda\int_0^\infty te^{-\lambda t}dt=\frac{1}{\lambda}. \qquad (2.23)$$

Получили ожидаемый результат: если λ среднее число событий за единицу времени, то средний промежуток времени между событиями равен $1/\lambda$.

Для дисперсии получим

$$D[T]=\int_0^\infty t^2\,f(t)dt-\frac{1}{\lambda^2}=\lambda\int_0^\infty t^2e^{-\lambda t}dt-\frac{1}{\lambda^2}=\frac{1}{\lambda^2}. \qquad (2.23)$$

4. *Равномерное распределение.* Равномерным распределением называется распределение для которого плотность равна постоянной величине на некотором промежутке $[a,b]$ и нулю вне этого промежутка:

$$f(x)=\begin{cases}0, & x\in(-\infty,a)\\ \dfrac{1}{b-a}, & x\in[a,b]\\ 0, & x\in(b,\infty)\end{cases}. \qquad (2.24)$$

Значение $f(x)$ на $[a,b]$ соответствует требованию: площадь под графиком плотности распределения должна быть равна единице.

Функцию распределения найдем в соответствии с (2.5). Внутри промежутка $[a,b]$ $F(x)$ определяется как

$$F(x)=\int_a^x\frac{1}{b-a}dx=\frac{x-a}{b-a}.$$

Полное выражение для $F(x)$ имеет вид

$$F(x)=\begin{cases}0, & x\in(-\infty,a)\\ \dfrac{x-a}{b-a}, & x\in[a,b]\\ 1, & x\in(b,\infty)\end{cases}. \qquad (2.25)$$

Несложно найти математическое ожидание и дисперсию. В силу того, что плотность распределения равна нулю вне промежутка $[a,b]$ интегралы по всей числовой прямой становятся интегралами по конечному промежутку:

$$M[X]=\int_a^b x\frac{1}{b-a}dx=\frac{a+b}{2}, \qquad (2.26)$$

$$D[X]=\int_a^b x^2\frac{1}{b-a}dx-\frac{(a+b)^2}{4}=\frac{(b-a)^2}{12}. \qquad (2.27)$$

Пример.

В конце пункта 2.1 были найдены функция распределения и плотность распределения случайной величины X – расстояние от пункта A до точки обрыва телефонной линии. Видим, что величина X имеет равномерное распределение с параметрами $a=0$, $b=L$. В соответствии с (2.26) и (2.27):

$$M[X]=\frac{L}{2},\ D[X]=\frac{L^2}{12},\ \sigma_X=\frac{L}{2\sqrt{3}}.$$

5. *Нормальное распределение* (*закон Гаусса*). Непрерывная случайная величина X распределена по нормальному закону (закону Гаусса) с параметрами a, σ если плотность распределения имеет вид

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-a)^2}{2\sigma^2}}. \tag{2.28}$$

Случайную величину с нормальным распределением будем называть нормальной случайной величиной.

Чтобы найти математическое ожидание и дисперсию нам потребуется интеграл Эйлера–Пуассона*:

$$\int_0^{\infty} e^{-x^2}dx=\frac{\sqrt{\pi}}{2}.$$

В силу четности подынтегральной функции

$$\int_{-\infty}^{\infty} e^{-x^2}dx=\sqrt{\pi}.$$

Используя этот результат несложно найти

$$M[X]=\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} xe^{-\frac{(x-a)^2}{2\sigma^2}}dx=a, \tag{2.29}$$

$$D[X]=\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} x^2e^{-\frac{(x-a)^2}{2\sigma^2}}dx-a^2=\sigma^2. \tag{2.30}$$

Таким образом, параметры a, σ в плотности нормального распределения (2.28) являются математическим ожиданием и средним квадратическим отклонением.

* Простое вычисление этого интеграла можно найти, например, в книге Фихтенгольц Г.М. Курс дифференциального и интегрального исчисления. – В 3-х т. Т. 3. – 11-е изд.Санкт-Петербург: Лань, 2020.

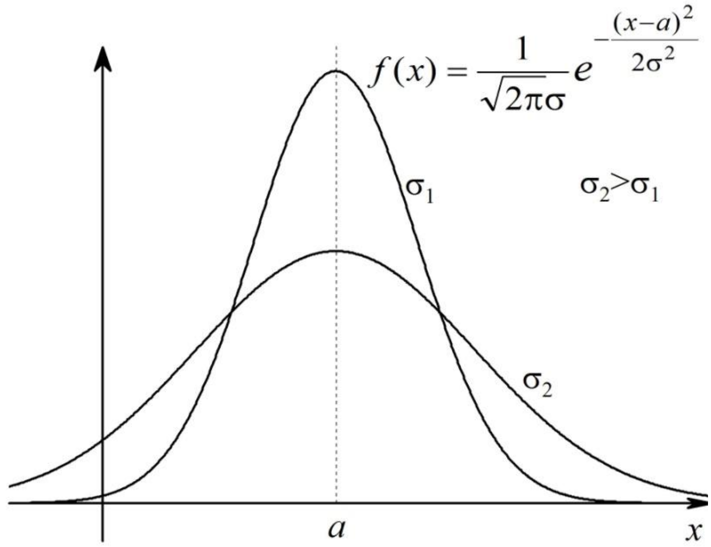


Рис.2.5.

Графики плотности нормального распределения, называемый кривой Гаусса или гауссоидой, представлены на рис. 2.5 для двух значений среднего квадратического отклонения.

В соответствии с (2.5) функция распределения нормальной случайной величины с математическим ожиданием m_x и средним квадратическим отклонением σ_x может быть найдена как

$$F(x)=\frac{1}{\sqrt{2\pi}\sigma_x}\int\limits_{-\infty}^{x}e^{-\frac{(t-m_x)^2}{2\sigma_x^2}}dt.$$

Интеграл в данном выражении не выражается через элементарные функции. Удобно ввести функцию

$$\Phi^*(x)=\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{x}e^{-t^2/2}dt, \qquad (2.31)$$

которая представляет собой нормальную функцию распределения с нулевым математическим ожиданием и единичной дисперсией. Теперь выразим функцию распределения $F(x)$ через $\Phi^*(x)$:

$$F(x)=\Phi^*\left(\frac{x-m_x}{\sigma_x}\right).$$

Непосредственно из определения (2.31) следует, что функция $\Phi^*(x)$ обладает всеми свойствами функции распределения. Кроме того, используя интеграл Эйлера–Пуассона, несложно получить $\Phi^*(0)=1/2$.

Рассмотрим функцию

$$\Phi(x)=\frac{1}{\sqrt{2\pi}}\int\limits_{0}^{x}e^{-t^2/2}dt, \qquad (2.32)$$

которая называется функцией Лапласа. Функции $\Phi^*(x)$ и $\Phi(x)$ связаны простым соотношением

$$\Phi^*(x)=\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{x}e^{-t^2/2}dt=\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{0}e^{-t^2/2}dt+\frac{1}{\sqrt{2\pi}}\int\limits_{0}^{x}e^{-t^2/2}dt=\frac{1}{2}+\Phi(x). \qquad (2.33)$$

Функция Лапласа обладает свойством нечетности $\Phi(-x)=-\Phi(x)$, что, как правило, делает ее более удобной при решении задач.

Графики функций $\Phi^*(x)$, $\Phi(x)$ изображены на рис. 2.6.

Таблица функции Лапласа $\Phi(x)$ приведена в приложении 1.

В соответствии с (2.2) выразим вероятность попадания нормальной случайной величины в промежуток $[a,b)$ через функции $\Phi^*(x)$, $\Phi(x)$:

$$P(a\le X<b)=\Phi^*\left(\frac{b-m_x}{\sigma_x}\right)-\Phi^*\left(\frac{a-m_x}{\sigma_x}\right)=\Phi\left(\frac{b-m_x}{\sigma_x}\right)-\Phi\left(\frac{a-m_x}{\sigma_x}\right). \qquad (2.34)$$

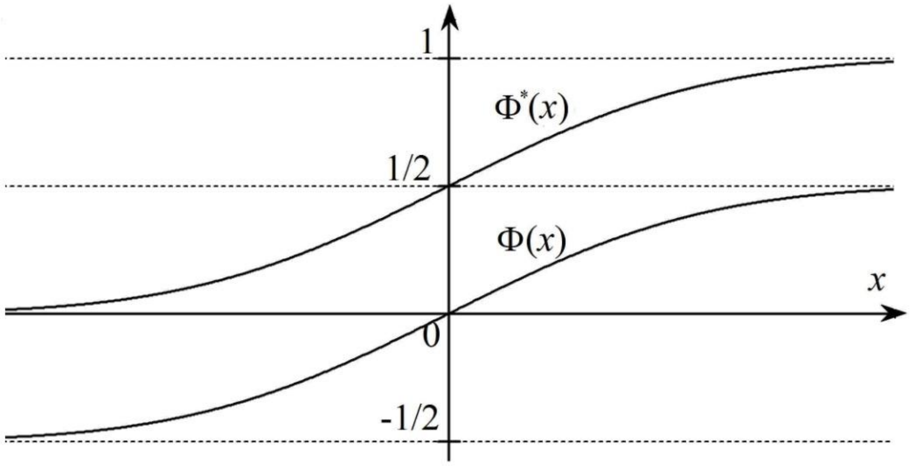


Рис. 2.6

Нормальное распределение играет исключительно важную роль в вероятностных и статистических задачах. Причина этого будет объяснена в следующей главе.

Примеры.

1) Правило «трех сигм». Найти вероятность того, что случайная величина X, распределенная нормальным образом, отклонится от среднего значения на величину не превышающую $3\sigma_x$.

В соответствии с (2.34)

$$P(|X-m_x|\le 3\sigma_x)=P(m_x-3\sigma_x\le X\le m_x+3\sigma_x)=\Phi\left(\frac{m_x+3\sigma_x-m_x}{\sigma_x}\right)-\Phi\left(\frac{m_x-3\sigma_x-m_x}{\sigma_x}\right)=$$

$$=\Phi(3)-\Phi(-3).$$

Используя свойство нечетности функции $\Phi(x)$ и определив по таблице приложения 1 значение $\Phi(3)$, получим

$$P(|X-m_x|\le 3\sigma_x)=2\Phi(3)=0{,}9973.$$

Соответственно, вероятность отклонения X от среднего на величину большую $3\sigma_x$ равна 0,0027, т.е. весьма мала.

2) Автомат изготавливает детали. Деталь считается годной, если отклонение ее длины X от проектной по абсолютной величине меньше 0,9 мм. Считая, что X распределена по нормальному закону с математическим ожиданием равным 0,05 мм и дисперсией 0,49 мм^2, найти процент годных деталей среди изготовленных.

Найдем вероятность того, что выбранная для проверки деталь окажется годной:

$$P(|X|<0{,}9)=P(-0{,}9<X<0{,}9)=\Phi\left(\frac{0{,}9-0{,}05}{0{,}7}\right)-\Phi\left(\frac{-0{,}9-0{,}05}{0{,}7}\right)\approx\Phi(1{,}21)-\Phi(-1{,}36)$$

$$=\Phi(1{,}21)+\Phi(1{,}36)=0{,}79994.$$

Таким образом, процент годных деталей приближенно составляет 80%.

3) Случайная величина X распределена нормально с математическим ожиданием 5,5 кг и дисперсией 1,21 кг^2. Найти интервал, симметричный относительно математического ожидания в который X попадет с вероятностью $\gamma=0{,}95$.

Интервал, симметричный относительно математического ожидания запишем как $(m_x-\varepsilon, m_x+\varepsilon)$.

Из (2.34) получим

$$P(m_x-\varepsilon<X<m_x+\varepsilon)=\Phi\left(\frac{m_x+\varepsilon-m_x}{\sigma_x}\right)-\Phi\left(\frac{m_x-\varepsilon-m_x}{\sigma_x}\right)=\Phi\left(\frac{\varepsilon}{\sigma_x}\right)-\Phi\left(\frac{-\varepsilon}{\sigma_x}\right)=2\Phi\left(\frac{\varepsilon}{\sigma_x}\right).$$

Так как по условию задачи данная вероятность известна, имеем уравнение для определения ε:

$$2\Phi\left(\frac{\varepsilon}{\sigma_x}\right)=\gamma,$$

$$\Phi\left(\frac{\varepsilon}{1{,}1}\right)=0{,}475.$$

По таблице приложения 1 для данного значения функции Лапласа определяем значение аргумента: $\varepsilon/1{,}1=1{,}96$ и $\varepsilon\approx2{,}16$.

Для интервала имеем $(3{,}34;\ 7{,}66)$.

2.4. Системы случайных величин

В вероятностных задачах часто приходится сталкиваться с ситуацией, когда результат опыта описывается не одной, а несколькими случайными величинами. Приведем примеры.

1. Стрелок стреляет по мишени. Если ввести систему координат на плоскости, то точка попадания представляет собой совокупность двух случайных величин X, Y.
2. С «продовольственной корзиной» связана система n случайных величин C_1, C_2,..., C_n – цены на входящие в нее продукты питания.
3. Успеваемость студента характеризуется системой n случайных величин $X_1, X_2,\ldots, X_n$ – оценок по n предметам.

Рассмотрим систему двух случайных величин X, Y.

Для дискретных случайных величин аналогом ряда распределения является матрица распределения, которая представляет собой таблицу, каждый элемент p_{ij} которой есть вероятность того, что X примет определенное значение x_i, а Y – значение y_j, что символически можно записать как

$$p_{ij}=P(\{X=x_i\}\cdot\{Y=y_j\}),$$

причем должно иметь место равенство $\sum_i\sum_j p_{ij}=1$.

Пусть X, Y дискретные случайные величины. В общем виде матрица распределения задается таблицей

y_j / x_i	y_1	y_2	…	y_j	…
x_1	p_{11}	p_{12}	…	p_{1j}	…
x_2	p_{21}	p_{22}	…	p_{2j}	…
…	….	…	…	…	…
x_i	p_{i1}	p_{i2}	…	p_{ij}	…
…	….	…	…	…	…

с неограниченным, вообще говоря, числом строк и/или столбцов.

Ряд распределения случайной величины *X*, т.е. совокупность вероятностей $p_i = P(X = x_i)$, может быть найден как

$$p_i = \sum_j p_{ij}.$$

Аналогично, ряд распределения случайной величины *Y*:

$$p_j = \sum_i p_{ij}.$$

Здесь и далее (в том случае, если это не вызывает недоразумений) вероятности, относящиеся к разным случайным величинам будем обозначать одной и той же буквой *p*, но с индексом *i* для *X* и *j* для *Y*.

Для математических ожиданий имеем:

$$m_x \equiv \mathrm{M}[X] = \sum_i x_i p_i = \sum_{ij} x_i p_{ij}, \quad m_y \equiv \mathrm{M}[Y] = \sum_j y_j p_j = \sum_{ij} y_j p_{ij}. \tag{2.35}$$

Моментом порядка *k*, *l* системы двух случайных величин называется математическое ожидание произведения X^k на Y^l:

$$\mu_{kl} = \mathrm{M}[X^k \cdot Y^l] = \sum_{i,j} x_i^k y_j^l p_{ij}. \tag{2.36}$$

Центральным моментом порядка *k*, *l* называется величина

$$\bar{\mu}_{kl} = \mathrm{M}[(X\text{-}m_x)^k \cdot (Y\text{-}m_y)^l] = \sum_{i,j} (x_i - m_x)^k (y_j - m_y)^l p_{ij}, \tag{2.37}$$

где m_x, m_y математические ожидания случайных величин *X, Y* соответственно.

Дисперсию *X* определяется как

$$D_x \equiv D[X] = \bar{\mu}_{20} = M\left[(X - m_x)^2\right]. \tag{2.38а}$$

$$D_x = \sum_{i,j} (x_i - m_x)^2 p_{ij} = \sum_{i,j} x_i^2 p_{ij} - 2m_x \sum_{i,j} x_i p_{ij} + m_x^2 \sum_{i,j} p_{ij} = M[X^2] - M^2[X]. \tag{2.38б}$$

Аналогично, для дисперсии *Y* имеем

$$D_y \equiv D[Y] = \bar{\mu}_{02} = M\left[(Y - m_y)^2\right], \tag{2.38в}$$

$$D_y = \sum_{i,j} (y_i - m_y)^2 p_{ij} = M[Y^2] - M^2[Y]. \tag{2.28г}$$

Корреляционным моментом или ковариацией называется величина

$$K_{xy} \equiv \bar{\mu}_{11} = \mathrm{M}\left[(X - m_x) \cdot (Y - m_y)\right]. \tag{2.39а}$$

Для дискретных случайных величин

$$K_{xy} = \sum_{i,j} (x_i - m_x) \cdot (y_j - m_y) p_{ij}. \tag{2.39б}$$

Пусть величины *X*, *Y* являются независимыми. Вероятность произведения независимых событий равна произведению вероятностей этих событий, поэтому

$$p_{ij} = P(\{X = x_i\} \cdot \{Y = y_j\}) = P(X = x_i) \cdot P(Y = y_j) = p_i \cdot p_j,$$

где $p_i \equiv P(X = x_i), p_j \equiv P(Y = y_j)$.

Дадим здесь некоторые свойства математического ожидания и дисперсии системы двух случайных величин.

1. $$\mathrm{M}[X+Y] = \sum_{i,j}(x_i + y_j)p_{ij} = \sum_{i,j} x_i p_{ij} + \sum_{i,j} y_j p_{ij} = \mathrm{M}[X] + \mathrm{M}[Y],$$

математическое ожидание суммы случайных величин равно сумме их математических ожиданий.

Пусть далее X, Y *независимые* случайные величины

2. $$\mathrm{M}[X \cdot Y] = \sum_{i,j} x_i y_j p_i p_j = \sum_i x_i p_i \cdot \sum_j y_j p_j = \mathrm{M}[X] \cdot \mathrm{M}[Y];$$

3. $$\mathrm{D}[X+Y] = \mathrm{M}\left[(X+Y)^2\right] - \mathrm{M}^2[X+Y] = \mathrm{M}\left[X^2\right] + \mathrm{M}\left[Y^2\right] + 2\,\mathrm{M}[X] \cdot \mathrm{M}[Y] -$$
$$- (\mathrm{M}[X] + \mathrm{M}[Y])^2 = \mathrm{M}\left[X^2\right] - \mathrm{M}^2[X] + \mathrm{M}\left[Y^2\right] - \mathrm{M}^2[Y] = \mathrm{D}[X] + \mathrm{D}[Y].$$

Таким образом, дисперсия суммы *независимых* случайных величин равна сумме их дисперсий.

Рассмотренные здесь свойства справедливы не только для двух, но для любого числа случайных величин.

Рассмотрим подробнее корреляционный момент K_{xy}

$$K_{xy} = \mathrm{M}\left[(X - m_x) \cdot (Y - m_y)\right] = \mathrm{M}\left[X \cdot Y - m_x Y - m_y X + m_x m_y\right] =$$
$$= \mathrm{M}[X \cdot Y] - m_x \mathrm{M}[Y] - m_y \mathrm{M}[X] + m_x m_y = \mathrm{M}[X \cdot Y] - m_x m_y - m_x m_y + m_x m_y =$$
$$= \mathrm{M}[X \cdot Y] - \mathrm{M}[X] \cdot \mathrm{M}[Y].$$

Таким образом, корреляционный момент может быть найден не только по формуле (2.39б), но и как

$$K_{xy} = \sum_{i,j} x_i y_j p_{ij} - m_x m_y. \qquad (2.39в)$$

Выражения (2.39б), (2.39в) совершенно равнозначны, но при практических расчетах, как правило удобнее пользоваться формулой (2.39в).

Для независимых СВ

$\mathrm{M}[X \cdot Y] = \mathrm{M}[X] \cdot \mathrm{M}[Y]$, следовательно, $K_{xy} = \mathrm{M}[X] \cdot \mathrm{M}[Y] - \mathrm{M}[X] \cdot \mathrm{M}[Y] = 0$.

Пусть теперь СВ *X*, *Y* связаны *очень сильной*, именно, *линейной* зависимостью: $Y = aX + b$. Тогда

$$K_{xy} = \mathrm{M}[X(aX+b)] - \mathrm{M}[X] \cdot \mathrm{M}[aX+b] = \mathrm{M}\left[aX^2 + bX\right] - \mathrm{M}[X] \cdot (a\,\mathrm{M}[X] + b) =$$
$$= a\,\mathrm{M}\left[X^2\right] + b\,\mathrm{M}[X] - a\,\mathrm{M}^2[X] - b\,\mathrm{M}[X] = a\left(\mathrm{M}\left[X^2\right] - \mathrm{M}^2[X]\right) = a\,\mathrm{D}[X].$$

Как характеристику того, насколько сильна корреляционная связь между случайными величинами вместо K_{xy} используют коэффициент корреляции r_{xy}:

$$r_{xy} = \frac{K_{xy}}{\sigma_x \sigma_y}, \qquad (2.40)$$

где σ_x, σ_y средние квадратичные отклонения величин X, Y: $\sigma_x = \sqrt{D[X]}$, $\sigma_y = \sqrt{D[Y]}$. Очевидно, для независимых случайных величин $r_{xy} = 0$. Если же для случайных величин имеет место сильная, т.е. линейная зависимость, то

$$r_{xy} = \frac{a\sigma_x^2}{\sigma_x\sqrt{a^2\sigma_x^2}} = \frac{a}{\sqrt{a^2}} = \frac{a}{|a|} = \begin{cases} +1, a \geq 0 \\ -1, a < 0 \end{cases}.$$

Таким образом, если $|r_{xy}|$ близок к единице, говорят о сильной корреляционной зависимости случайных величин. И наоборот, близость $|r_{xy}|$ к нулю говорит о слабой корреляционной связи*. Более точные количественные оценки, позволяющие говорить о сильной (слабой) корреляционной связи будут даны ниже.

Подчеркнем здесь следующее обстоятельство. Для случайных величин не следует смешивать понятия «некоррелированность» и «независимость». Если случайные величины X, Y независимы ($P(\{X = x\} \cdot \{Y = y\}) = P(X = x) \cdot P(Y = y)$), то они и некоррелированны, то есть $r_{xy} = 0$. Но обратное, вообще говоря, неверно: если случайные величины некоррелированны, то это вовсе не означает, что они независимы. Это может означать также, что зависимость между ними существенно нелинейная. Можно сказать, что отсутствие корреляции (r_{xy}=0) – необходимое, но не достаточное условие независимости.

Пример. По многолетним данным получена следующая матрица распределения случайных величин X – посещаемость занятий по некоторой дисциплине, Y – успеваемость по данной дисциплине

Таблица 2.1

Посещаемость, %	X	Успеваемость, Y			
		неуд(**2**)	удовл(**3**)	Хорошо(**4**)	Отлич(**5**)
0-20	10	0,08	0,05	0	0,01
20-40	30	0,045	0,08	0,07	0
40-60	50	0,01	0,08	0,085	0,015
60-80	70	0,005	0,04	0,12	0,065
80-100	90	0	0,085	0,1	0,06

Найти m_x, m_y, σ_x, σ_y, r_{xy}.

Прежде всего, заметим, что случайной величине X мы присвоим значения равные серединам интервалов, что является обычным приемом в математической статистике. Эти значения записаны во втором столбце таблицы.

Теперь

$$m_x = 10 \cdot (0{,}08 + 0{,}05 + 0 + 0{,}01) + 30 \cdot (0{,}045 + 0{,}08 + 0{,}07 + 0) +$$
$$+ 50 \cdot (0{,}01 + 0{,}08 + 0{,}085 + 0{,}015) + 70 \cdot (0{,}005 + 0{,}04 + 0{,}12 + 0{,}065) +$$
$$+ 90 \cdot (0 + 0{,}085 + 0{,}1 + 0{,}06);$$

$$m_x = 54{,}9.$$

* Следует иметь в виду, что здесь под зависимостью случайных величин подразумевается функциональная зависимость, которая вовсе не предполагает наличие причинно-следственной связи между ними.

$$m_y = 2\cdot(0{,}08+0{,}045+0{,}01+0{,}005+0)+3\cdot(0{,}05+0{,}08+0{,}08+0{,}04+0{,}085)+$$
$$+4\cdot(0+0{,}07+0{,}085+0{,}12+0{,}1)+5\cdot(0{,}01+0+0{,}015+0{,}065+0{,}06);$$
$$m_y = 3{,}535.$$
$$\sigma_x^2 = 10^2\cdot(0{,}08+0{,}05+0+0{,}01)+30^2\cdot(0{,}045+0{,}08+0{,}07+0)+$$
$$+50^2\cdot(0{,}01+0{,}08+0{,}085+0{,}015)+70^2\cdot(0{,}005+0{,}04+0{,}12+0{,}065)+$$
$$+90^2\cdot(0+0{,}085+0{,}1+0{,}06)-54{,}9^2;$$
$$\sigma_x^2 = 761{,}99.$$
$$m_y = 2^2\cdot(0{,}08+0{,}045+0{,}01+0{,}005+0)+3^2\cdot(0{,}05+0{,}08+0{,}08+0{,}04+0{,}085)+$$
$$+4^2\cdot(0+0{,}07+0{,}085+0{,}012+0{,}1)+5^2\cdot(0{,}01+0+0{,}015+0{,}065+0{,}06)-3{,}535^2;$$
$$\sigma_y^2 = 0{,}829.$$
$$K_{xy} = 10\cdot2\cdot0{,}08+10\cdot3\cdot0{,}05+10\cdot4\cdot0+10\cdot5\cdot0{,}01+$$
$$+30\cdot2\cdot0{,}045+30\cdot3\cdot0{,}08+30\cdot4\cdot0{,}07+30\cdot5\cdot0+$$
$$+50\cdot2\cdot0{,}01+50\cdot3\cdot0{,}08+50\cdot4\cdot0{,}085+50\cdot5\cdot0{,}015+$$
$$+70\cdot2\cdot0{,}005+70\cdot3\cdot0{,}04+70\cdot4\cdot0{,}12+70\cdot5\cdot0{,}065+$$
$$+90\cdot2\cdot0+90\cdot3\cdot0{,}085+90\cdot4\cdot0{,}1+90\cdot5\cdot0{,}06-54{,}9\cdot3{,}535;$$
$$K_{xy} \approx 12{,}979.$$
$$r_{xy} \approx \frac{12{,}979}{\sqrt{761{,}99}\cdot\sqrt{0{,}829}} \approx 0{,}516.$$

Часто при расчетах таблицы подобные таблице 2.1 модифицируют. Именно, справа приписывают столбец, в которой записываю сумму вероятностей по строкам, в наших обозначениях это p_i. Внизу дописывают суммы вероятностей по столбцам, в наших обозначениях это p_j. Таким образом, имеем таблице 2.2

Таблица 2.2

Посещаемость %		Успеваемость				
		2	3	4	5	p_i
0-20	10	0,08	0,05	0	0,01	0,14
20-40	30	0,045	0,08	0,07	0	0,195
40-60	50	0,01	0,08	0,085	0,015	0,19
60-80	70	0,005	0,04	0,12	0,065	0,23
80-100	90	0	0,085	0,1	0,06	0,245
	p_j	0,14	0,335	0,375	0,15	

Таким образом, для m_x получим

$$m_x = 10\cdot0{,}14+30\cdot0{,}195+50\cdot0{,}19+70\cdot0{,}23+90\cdot0{,}245 = 54{,}9$$

и т.д.

А вообще, такого рода расчеты удобно выполнять в Excel.

Функция распределения $F(x,y)$, определяется как вероятность *совместного* выполнения двух неравенств: $X < x$ и $Y < y$. Таким образом, можно записать

$$F(x,y) = P(\{X < x\} \cdot \{Y < y\}). \tag{2.35}$$

Из самого определения функции $F(x,y)$ вытекают следующие ее свойства.

1. $\lim\limits_{x \to -\infty} F(x,y) = \lim\limits_{y \to -\infty} F(x,y) = \lim\limits_{x \to -\infty} \lim\limits_{y \to -\infty} F(x,y) = 0$,

2. $\lim\limits_{x \to +\infty} \lim\limits_{y \to +\infty} F(x,y) = 1$,

3. если $x_2 > x_1$ для данного *y*, то $F(x_2,y) \ge F(x_1,y)$, аналогично, если $y_2 > y_1$ для данного *x*, то $F(x,y_2) \ge F(x,y_1)$.

Если ввести систему координат, то каждой паре определенных значений величин *X, Y* будет соответствовать точка плоскости. Рассмотрим на плоскости *x*0*y* прямоугольник с координатами левого нижнего угла (a_1, b_1), правого верхнего – (a_2, b_2) – см. рис.2.7.

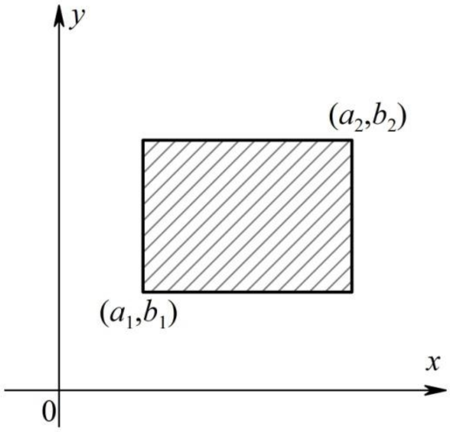


Рис. 2.7.

Так как

$$P(\{X < a_2\} \cdot \{Y < b_2\}) = P(\{X < a_1\} \cdot \{Y < b_2\}) + P(\{X < a_2\} \cdot \{Y < b_1\}) - P(\{X < a_1\} \cdot \{Y < b_1\}) +$$
$$+ P(\{a_1 < X < a_2\} \cdot \{b_1 < Y < b_2\})$$

то

$$P(\{a_1 < X < a_2\} \cdot \{b_1 < Y < b_2\}) = P(\{X < a_2\} \cdot \{Y < b_2\}) - P(\{X < a_1\} \cdot \{Y < b_2\}) -$$
$$- P(\{X < a_2\} \cdot \{Y < b_1\}) + P(\{X < a_1\} \cdot \{Y < b_1\}).$$

Теперь, непосредственно из определения функции распределения следует, что вероятность попадания случайной точки $M(X,Y)$ в данный прямоугольник равна

$$P(\{a_1 < X < a_2\} \cdot \{b_1 < Y < b_2\}) = F(a_2,b_2) - F(a_1,b_2) - F(a_2,b_1) + F(a_1,b_1). \tag{2.41}$$

Плотность распределения вероятности системы двух непрерывных случайных величин определяется как смешанная производная

$$f(x,y) = \frac{\partial^2 F(x,y)}{\partial x \partial y}. \tag{2.42}$$

Имеет место следующее свойство

$$\int\limits_{-\infty}^{\infty} \int\limits_{-\infty}^{\infty} f(x,y) = 1. \tag{2.43}$$

Для независимых случайных величин

$$F(x,y) = P(X < x) \cdot P(Y < y) = F_1(x) F_2(y),$$

где $F_1(x)$, $F_2(y)$ – функции распределения X, Y соответственно. Для плотности распределения в этом случае получим

$$f(x,y) = \frac{dF_1(x)}{dx} \cdot \frac{dF_2(y)}{dy} = f_1(x) f_2(y).$$

Математические ожидания непрерывных случайных величин определяются как

$$M[X] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f(x,y)\,dxdy, \quad M[Y] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y f(x,y)\,dxdy. \qquad (2.44)$$

Дисперсии X, Y по-прежнему определяются формулами (2.38а), (2.38в), а корреляционный момент формулой (2.39а).

Задачи для самостоятельного решения.

1) Вероятность появления события A в некотором испытании равна 1/4. Проводится десять независимых испытаний. Случайная величина X – число появлений A . Найти вероятность того, что X примет значение, не превышающее ее математическое ожидание.

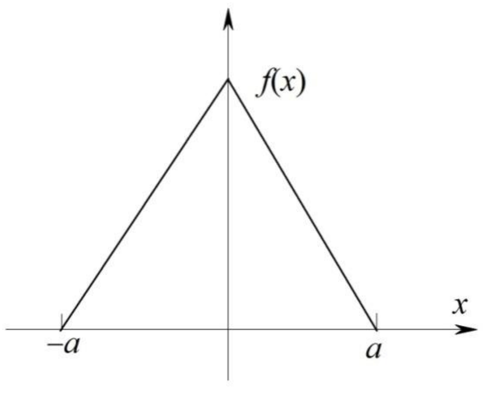
2) Поток заявок, поступающих на телефонную станцию, представляет собой пуассоновский (простейший) поток. Математическое ожидание числа вызовов за час равно 30. Найти вероятность того, что за минуту поступит не менее двух вызовов.

3) Число атак истребителей, которым может подвергнуться бомбардировщик над территорией противника, есть случайная величина, распределенная по закону Пуассона с математическим ожиданием a=3. Каждая атака с вероятностью p=0,4 заканчивается поражением бомбардировщика. Определить вероятность поражения бомбардировщика.

4) Случайная величина X распределена равномерно с математическим ожиданием равным –1 и дисперсией равной 3. Записать функцию распределения и плотность распределения данной случайной величины. Найти вероятность попадания случайной величины на интервал [0, 9).

5) Случайная величина T распределена по показательному закону. Найти вероятность того, что значение случайной величины T превысит ее удвоенное среднее значение.

6) Плотность распределения случайной величины X имеет вид, изображенный на рисунке



Найти математическое ожидание, дисперсию, среднее квадратическое отклонение данной случайной величины и вероятность попадания на промежуток $[a/2, 2a]$.

7) Найти вероятность того, что случайная величина X отклонится от своего математического ожидания на величину большую σ_x, если случайная величина распределена

а) равномерно;

б) по показательному закону;

в) нормально.

8) На двух станках-автоматах изготавливают заготовки, которые представляют собой прямоугольники со сторонами *x*, *y*. Автоматы работают независимо: сначала первый автомат вырезает полосу шириной *y*, затем второй разрезает ее на прямоугольники длины *x*. Из-за люфта режущих кромок станков неизбежны отклонения *X*, *Y* длины и ширины заготовки от проектных значений. Заготовка считается бракованной, если отклонение от проектного хотя бы одного из размеров превышает 2 мм. Рассматривая *X*, *Y* как нормальные случайные величины с математические ожидания равными нулю и дисперсиями равными D_x=0,0196 см^2, D_y=0,01 см^2 соответственно, найти процент брака.

9) Нормальная случайная величина X имеет математическое ожидание, равное 6 и дисперсию 0,25. Найти интервал, симметричный относительно математического ожидания, в который данная случайная величина попадает с вероятностью 0,99.

10) Найти коэффициент корреляции r_{xy} для матрицы из таблицы

Посещаемость %		Успеваемость			
		2	3	4	5
0-20	10	0,1	0,03	0,01	0
20-40	30	0,05	0,065	0,075	0
40-60	50	0,025	0,07	0,07	0,02
60-80	70	0	0,05	0,115	0,065
80-100	90	0,005	0,01	0,13	0,11

Ответы: 1) $7\left(\frac{3}{4}\right)^9 \approx 0{,}526$. 2) 0,0902. 3) $1-e^{-ap} \approx 0{,}699$.

4) $f(x)=\begin{cases} 0, & x \le -4, \\ 1/6, & -4 < x \le 2 \\ 0, & x > 2 \end{cases}$, $F(x)=\begin{cases} 0, & x \le -4, \\ (x+4)/6, & -4 < x \le 2 \\ 1, & x > 2 \end{cases}$, $P(0 \le X < 9) = 1/3$. 5) e^{-2}. 6) 0, $a^2/6$, $a/\sqrt{6}$, 1/4. 7) а) 0,423; б) $e^{-2} \approx 0{,}135$; в) 0,317. 8) Приближенно19,1%. 9) (4,71; 7,29). 10) $r_{xy} \approx 0{,}67$.

Глава 3. Предельные теоремы теории вероятностей

3.1. Закон больших чисел

Здесь нам потребуется такое понятие как *сходимость по вероятности*, которое вводится для последовательностей, члены которых являются случайными величинами. Простейшим примером такой последовательности является, например, частота появлений герба при подбрасывании монеты 10, 100, 1000, …раз. Для случайной последовательности невозможно дать определение предела, которое дается для обычной последовательности, члены которой зависят от номера n в соответствии с определенной строго детерминированной формулой. Для случайных последовательностей можно говорить только о сходимости вероятности отклонения членов последовательности от некоторого неслучайного значения.

Пусть имеется последовательность случайных величин $X_1, X_2, X_3, \ldots, X_n$. Говорят, что данная последовательность сходится по вероятности (при $n \to \infty$) к неслучайной величине ξ, если $\lim\limits_{n\to\infty} P(|X_n - \xi| < \varepsilon) = 1$, где ε любое положительное *сколь угодно малое* число.

Далее сходимость по вероятности будем обозначать $X_n \xrightarrow{P} \xi$.

Закон больших чисел (*первая теорема Чебышева*). Пусть X случайная величина с математическим ожиданием m_x. Пусть в результате n *независимых* опытов получены следующие значения данной случайной величины: $x_1,\ x_2,\ x_3, \ldots, x_n$. Составим среднее арифметическое* полученных значений

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} x_i .$$

При $n \to \infty$ среднее арифметическое сходится по вероятности к математическому ожиданию m_x: $\overline{X}_n \xrightarrow{P} m_x$.

Доказательство.

Прежде всего, получим неравенство, известное как неравенство Чебышева. Пусть X дискретная случайная величина. Упорядочим ее возможные значения от меньших значений к большим: $x_1 < x_2 < x_3 < \ldots$. Рассмотрим вероятность $P(|X - m_x| \geq \varepsilon)$. Обозначим как j_1 последний индекс, для которого имеет место неравенство $x_i \leq m_x - \varepsilon$, j_2 как первый индекс для которого $x_i \geq m_x + \varepsilon$. Тогда

$$P(|X - m_x| \geq \varepsilon) = \sum_{i \leq j_1} p_i + \sum_{i \geq j_2} p_i . \qquad (*)$$

* Заметим, что среднее арифметическое следует рассматривать как случайную величину, так как наблюденные значения случайной величины могут меняться непредсказуемым образом при повторении измерений.

$$D_x = \sum_i (x_i - m_x)^2 p_i = \sum_{i \le j_1} (x_i - m_x)^2 p_i + \sum_{i=j_1+1}^{j_2-1} (x_i - m_x)^2 p_i + \sum_{i \ge j_2} (x_i - m_x)^2 p_i \, ,$$

$$D_x \ge \sum_{i \le j_1} (x_i - m_x)^2 p_i + \sum_{i \ge j_2} (x_i - m_x)^2 p_i \, .$$

В первой сумме $x \le m_x - \varepsilon$, то есть, $x - m_x \le -\varepsilon$. Во второй сумме $x \ge m_x + \varepsilon$ и $x - m_x \ge \varepsilon$. Два неравенства можно объединить в одно: $(x - m_x)^2 \ge \varepsilon^2$

Таким образом, для дисперсии имеет место неравенство

$$D_x \ge \varepsilon^2 \left(\sum_{i \le j_1} p_i + \sum_{i \ge j_2} p_i \right). \qquad (**)$$

Из неравенств (*), (**) следует неравенство Чебышева

$$P(| X - m_x |\ge \varepsilon) \le \frac{D_x}{\varepsilon^2} \, . \qquad (3.1)$$

Аналогичным образом, с интегралами вместо сумм, данное неравенство можно доказать и для непрерывных случайных величин.

Найдем математическое ожидание и дисперсию величины $\overline{X}_n$. Следует иметь в виду, что наблюденные значения x_1, x_2, $x_3, \dots, x_n$ сами являются случайными величинами: несколько наблюдателей, выполняя *n* измерений одной и той же величины *X*, будут получать, вообще говоря, разные наборы значений. Так как все эти значения являются реализациями одной и той же случайной величины, их математические ожидания одинаковы и равны m_x. Используя свойства математического ожидания, получим

$$M[\overline{X}_n] = M\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] = \frac{1}{n}\sum_{i=1}^{n} M[x_i] = \frac{1}{n} \cdot n \cdot m_x = m_x \, .$$

Так как дисперсия суммы независимых случайных величин равна сумме их дисперсий, то

$$D[\overline{X}_n] = D\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} D[x_i] = \frac{1}{n^2} \cdot n \cdot m_x = \frac{m_x}{n} \, .$$

Теперь применим к $\overline{X}_n$ неравенство Чебышева:

$$P(| \overline{X}_n - m_x |\ge \varepsilon) \le \frac{D_x}{n\varepsilon^2}$$

Видим, что при $n \to \infty$ $P(| \overline{X}_n - m_x |\ge \varepsilon) \to 0$, следовательно, $P(| \overline{X}_n - m_x |< \varepsilon) \to 1$, что и доказывает теорему.

Теорема Бернулли (следствие закона больших чисел). Пусть проведено n независимых опытов, в каждом из которых событие А появляется с вероятностью p . При $n \to \infty$ частота ν_n события А сходится по вероятности к p.

Напомним здесь, что частотой ν_n события А в n опытах называется отношение числа опытов в которых событие А произошло к общему числу опытов.

Доказательство.

Рассмотрим случайные величины $X_i, i = 1, 2, 3, ..., n$:

$$X_i = \begin{cases} 1, & \text{если в } i\text{–м опыте событие А произошло,} \\ 0, & \text{если в } i\text{–м опыте А не произошло.} \end{cases} \tag{3.2}$$

Распределение случайной величины X_i называется распределением Бернулли.

Математическое ожидание определяется элементарно

$M[X_i] = 1 \cdot p + 0 \cdot (1 - p) = p$.

Число появлений события А в n опытах равно сумме всех наблюденных значений x_i величин X_i , следовательно частота равна среднему арифметическому этих величин:

$$\nu_n = \frac{1}{n} \sum_{i=1}^{n} x_i \ .$$

Тогда из закона больших чисел немедленно следует $\nu_n \xrightarrow{P} p$.

3.2. Центральная предельная теорема (ЦПТ)

Мы дадим здесь данную теорему в следующей формулировке. Пусть имеется n независимых случайных величин $X_1, X_2, X_3, \ldots, X_n$, имеющих нулевое математическое ожидание и единичную дисперсию. Составим случайную величину

$$Y_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i$$

При неограниченном увеличении числа слагаемых n распределение случайной величины Y_n неограниченно приближается к нормальному с нулевым математическим ожиданием и единичной дисперсией, т.е. плотность распределения величины Y_n при достаточно больших n приближенно равна

$$f(y) = \frac{1}{\sqrt{2\pi}} e^{-y^2/2}.$$

Может показаться, что условия, наложенные на математическое ожидание и дисперсию, значительно ограничивают общность теоремы, но это не так. Пусть X произвольная случайная величина с математическим ожиданием m_x и дисперсией $\mathrm{D}[X] = \sigma_x^2$. Из X всегда можно сделать случайную величину с нулевым математическим ожиданием, вычитая m_x . Если далее, разделить получившуюся величину на σ_x, получим величину с единичной дисперсией. То есть случайная величина $(X - m_x)/\sigma_x$ имеет нулевое математическое ожидание и единичную дисперсию. Требования, наложенные на случайные величины в приведенной здесь формулировке ЦПТ необходимы только для того, чтобы упростить ее применения.

Смысл теоремы заключается в том, что распределение суммы независимых случайных величин неограниченно приближается к нормальному при неограниченном увеличении числа слагаемых.

Теорему даем здесь без доказательства.

Теорема Муавра-Лапласа. Пусть производится n независимых опытов, в каждом из которых событие А появляется с вероятностью p. Рассмотрим случайную величину X – число появлений события А в n независимых опытах. Имеет место следующее приближенное равенство

$$P\left(\alpha \le \frac{X-np}{\sqrt{npq}} < \beta\right) \approx \Phi(\beta)-\Phi(\alpha), \tag{3.3}$$

где $q = 1 - p$. Данное приближенное равенство выполняется тем точнее, чем больше n.

Данная теорема является исторически первой доказанной формой ЦПТ. Но мы приведем здесь ее доказательство, опираясь на ЦПТ.

Доказательство.

Рассмотрим случайные величины X_i, $i = 1, 2, 3, \ldots, n$, распределенные по закону (3.2). Так как

$$M[X_i] = p, \; D[X_i] = 1^2 \cdot p + 0 \cdot q - p^2 = p - p^2 = pq,$$

то случайная величина $\dfrac{X_i - p}{\sqrt{pq}}$ имеет нулевое математическое ожидание и единичную дисперсию. Тогда согласно ЦПТ распределением случайной величины

$$Y_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{X_i - p}{\sqrt{pq}}$$

близко к нормальному с нулевым математическим ожиданием и единичной дисперсией. Но

$$Y_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{X_i - p}{\sqrt{pq}} = \frac{\sum_{i=1}^{n}(X_i - p)}{\sqrt{npq}} = \frac{\sum_{i=1}^{n}X_i - np}{\sqrt{npq}} = \frac{X - np}{\sqrt{npq}},$$

и, следовательно,

$$P\left(\alpha \le \frac{X-np}{\sqrt{npq}} < \beta\right) \approx \Phi(\beta)-\Phi(\alpha) \; .$$

При практических расчетах приведенное равенство не очень удобно. Видоизменим его следующим образом.

$$\alpha \le \frac{X-np}{\sqrt{npq}} < \beta \Rightarrow \alpha\sqrt{npq} \le X-np < \beta\sqrt{npq} \Rightarrow np+\alpha\sqrt{npq} \le X < np+\beta\sqrt{npq}\,.$$

Обозначим $a \equiv np+\alpha\sqrt{npq}, b \equiv np+\beta\sqrt{npq}$. Тогда

$$P(a \le X < b) \approx \Phi^*\left(\frac{b-np}{\sqrt{npq}}\right)-\Phi^*\left(\frac{a-np}{\sqrt{npq}}\right)=\Phi\left(\frac{b-np}{\sqrt{npq}}\right)-\Phi\left(\frac{a-np}{\sqrt{npq}}\right)\,. \tag{3.4}$$

Заметим здесь, что точным распределение случайной величины X является распределение Бернулли, математическое ожидание которого np, а дисперсия npq.

Пример. Игральный кубик бросается $n=1000$ раз. Случайная величина X – число появлений шестерки. Найти интервал, симметричный относительно математического ожидания, в который данная случайная величина попадает с вероятностью $\gamma = 0.99$.

Данная случайная величина описывается биноминальным распределением. Однако использовать это распределение для вычислений в данном случае практически невозможно. Применим теорему Муавра-Лапласа.

$M[X]=np$, где $p=1/6$ – вероятность появления шестерки в отдельно взятом испытании. Интервал симметричный относительно математического ожидания запишем в виде $(np-\varepsilon, np+\varepsilon)$. Таким образом, необходимо найти единственную величину ε.

$$P(np-\varepsilon \le X < np+\varepsilon) \approx \Phi\left(\frac{np+\varepsilon-np}{\sqrt{npq}}\right)-\Phi\left(\frac{np-\varepsilon-np}{\sqrt{npq}}\right)=$$

$$=\Phi\left(\frac{\varepsilon}{\sqrt{npq}}\right)-\Phi\left(-\frac{\varepsilon}{\sqrt{npq}}\right)=2\Phi\left(\frac{\varepsilon}{\sqrt{npq}}\right).$$

Здесь мы воспользовались свойством нечетности функции $\Phi(x)$. Величину ε найдем из уравнения

$$2\Phi\left(\frac{\varepsilon}{\sqrt{npq}}\right)=0.99,\ \Phi\left(\frac{\varepsilon}{\sqrt{npq}}\right)=0.495\,.$$

Воспользовавшись таблицей функции Лапласа, получим

$$\frac{\varepsilon}{\sqrt{npq}} \approx 2.58,\ \varepsilon \approx 2.58\sqrt{1000\cdot\frac{1}{6}\cdot\frac{5}{6}} \approx 30.41$$

С учетом того, что величина X целая, получим окончательный ответ: $(136{,}198)$.

С достаточно хорошей точностью можно заменить биноминальное распределение нормальным, если выполняются условия

$$np-3\sqrt{npq} > 0, np+3\sqrt{npq} < n\,.$$

Используя теорему Муавра-Лапласа, можно найти приближенно вероятность того, что случайная величина X, имеющая биноминальное распределение примет определенное значение m:

$$P(X=m)\approx\Phi\left(\frac{m+1-np}{\sqrt{npq}}\right)-\Phi\left(\frac{m-np}{\sqrt{npq}}\right)=\Phi\left(\frac{m-np}{\sqrt{npq}}+\frac{1}{\sqrt{npq}}\right)-\Phi\left(\frac{m-np}{\sqrt{npq}}\right).$$

Найдем приближенное выражение для разности функций Лапласа

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

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

Таким образом

$$P(X=m)\approx\frac{1}{\sqrt{2\pi npq}}e^{-(m-np)^2/2npq}.$$

Полученное соотношение называют точечной теоремой Муавра–Лапласа.

Пример. Пусть монета бросается 10 раз: n=10. Найдем точное и приближенное вероятности того, что герб появится ровно пять раз.

Точный ответ дает, разумеется, биноминальное распределение $P_{10}(5)=C_{10}^{5}\cdot\frac{1}{2^{10}}\approx 0{,}246$

$$np-3\sqrt{npq}=5-3\sqrt{10\cdot\frac{1}{4}}=5-3\sqrt{2{,}5}>0,$$

$$np+3\sqrt{npq}=5+3\sqrt{2{,}5}<10$$

Приближенный ответ

$$P(X=5)\approx\frac{1}{\sqrt{2\pi\cdot 10\cdot 1/4}}e^{-(5-5)^2/(2\cdot 10\cdot 1/4)}=\frac{1}{\sqrt{2\pi\cdot 2.5}}\approx 0.252.$$

Видим, что существенной разницы между точным и приближенным ответами нет.

Задачи для самостоятельного решения.

1) Производится n =1000 независимых испытаний, в каждом из которых событие A наблюдается с вероятностью 1/2. Случайная величина X – число появлений A. Найти интервал, симметричный относительно математического ожидания, в который X попадет с вероятностью 0,95.

2) Цех производит шарики для подшипников. За смену производится n=10000 шариков. Вероятность того, что шарик окажется дефектным равна 0,05. Продукция проходит контроль, дефектные шарики бракуются и ссыпаются в специальный бункер. Определить на какое количество шариков должен быть рассчитан бункер, чтобы с вероятностью 0,99 он не оказался переполненным. *Указание*: количество бракованных шариков за смену есть

случайная величина X. Искомая величина m может быть найдена из уравнения $P(0 \le X \le m) = 0{,}95$, в котором вероятность P следует записать, используя соотношение (3.4).

3) Автомат изготавливает детали. Автомат не требует наладки при условии, что процент брака (отклонение детали от проектного размера) не превышает 0,5%. Изготовлено 4000 деталей. Определить промежуток, в который с вероятностью 0,99 должно попасть число годных деталей при условии, что автомат не требует наладки. Предполагается, что детали оказываются бракованными независимо друг от друга и автомат не дает систематической ошибки (математическое ожидание отклонения детали равно нулю).

Ответы. 1) [469, 531]. 2) >551. 3) [3968, 4000].

Глава 4. Основные задачи математической статистики.

4.1. Эмпирические ряды распределения и их графические представления.

Пусть имеется совокупность (иначе множество) *всех возможных* значений некоторого признака, характеризующего какую-то величину. Например, величина месячного дохода *для каждого* жителя некоторого региона, вес колоса для *всех* побегов пшеницы на опытном поле, отклонение от стандартного размера для *очень крупной* партии деталей и т.д. В статистике такую совокупность называют *генеральной совокупностью.* В силу того, что генеральная совокупность очень велика, практически невозможно получить значение признака для каждого экземпляра совокупности. Поэтому из генеральной совокупности выбирают существенно меньшее число экземпляров, именно такое, что получить значение признака для каждого экземпляра практически возможно. В статистике это называется *выборкой*. Изучая выборку, пытаются сделать выводы о свойствах всей совокупности. Множество всех возможных значений признака генеральной совокупности соответствует множеству всех возможных значений случайной величины в теории вероятностей. Поэтому понятие генеральной совокупности аналогично понятию случайной величины. Что касается выборки, то в терминах теории вероятностей это множество значений $x_1, x_2, \ldots, x_n$ случайной величины X, полученное в результате ограниченного числа опытов.

Статистическим или вариационным рядом называют таблицу в первой строке которой стоят определенные значения изучаемого признака, (как правило, ранжированные, т.е. упорядоченные по возрастанию), во второй – число повторений этого значения.

В статистике число повторений данного значения называют частотой, число повторений, деленное на общее число опытов (измерений) – относительной частотой или частостью. Заметим здесь, что в теории вероятностей обычно относительную частоту называют просто частотой, а термин «относительная частота» в таком случае вовсе не используют.

Пример. В некотором ВУЗе получены данные по успеваемости по математике для выборки из ста студентов. Эти данные представлены в виде вариационного ряда (табл. 4.1). Частота здесь – это число студентов, имеющих данную оценку. Сумма всех частот равна объему выборки, т.е. 100 в данном примере. Такой вариационный ряд называется дискретным или точечным.

Таблица 4.1

Оценка x_i	2 («неудовлетворительно»)	3 («удовлетворительно»)	4 («хорошо»)	5 («отлично»)
Частота, m_i	13	36	28	23

Если число возможных значений признака велико, построение дискретного вариационного ряда в виде таблицы 4.1 становится затруднительно, кроме того ряд потеряет наглядность. В этом случае весь диапазон значений разбивается на промежутки или разряды. Такое разбиение имеет еще и то преимущество, что сглаживает статистически незначащие выбросы возможных значений признака. Число промежутков должно быть таково, чтобы с одной стороны, с вариационным рядом было удобно работать, с другой стороны – оно не должно быть слишком мало, чтобы не потерять статистически значимые вариации признака. Часто для определения числа разрядов используют полуэмпирическое правило Стерджеса

$$2^{k-1} = n \text{ или } k = \log_2 n + 1, \qquad (4.1)$$

где k – число разрядов, n – объем выборки. Величина интервала Δx определяется как $\Delta x = (x_{\max} - x_{\min})/k$.

Разбиение на разряды не имеет смысла, если мы имеем дело с наблюденными значениями *дискретной* случайной величины, причем множество возможных значений невелико. Разбиение является, фактически, необходимой процедурой, если имеется достаточно большая выборка значений (по крайней мере, несколько десятков) *непрерывной* случайной величины.

Пример статистического ряда с группировкой по разрядам приведен в таблице 4.2. В первой строке таблицы содержится уровень месячного дохода для жителей некоторого региона, по второй строке – число жителей с данным уровнем дохода для выборки с объемом $n = 1000$.

Таблица 4.2

Доход x_i тыс.руб.	0–5	5–10	10–15	15–20	20-25	25-30	30-35
Число жителей m_i	58	96	140	220	298	157	31

При группировке наблюденных значений признака по разрядам может возникнуть вопрос: к какому разряду отнести значение, находящееся в точности на их границе. Здесь поступают по-разному. Например, левую границу включают в данный разряд, а правую нет. Или считают, что данное значение принадлежит обоим разрядам и добавляют к значению соответствующих частот число $1/2$.

Графическое изображение дискретного вариационного ряда называется полигоном. Абсцисса некоторой точки полигона представляет собой значение признака, ордината – соответствующую относительную частоту $\nu_i = m_i/n$. Точки обычно соединяют прямыми линиями. Полигон, соответствующий таблице 4.1 изображен на рис. 4.1.

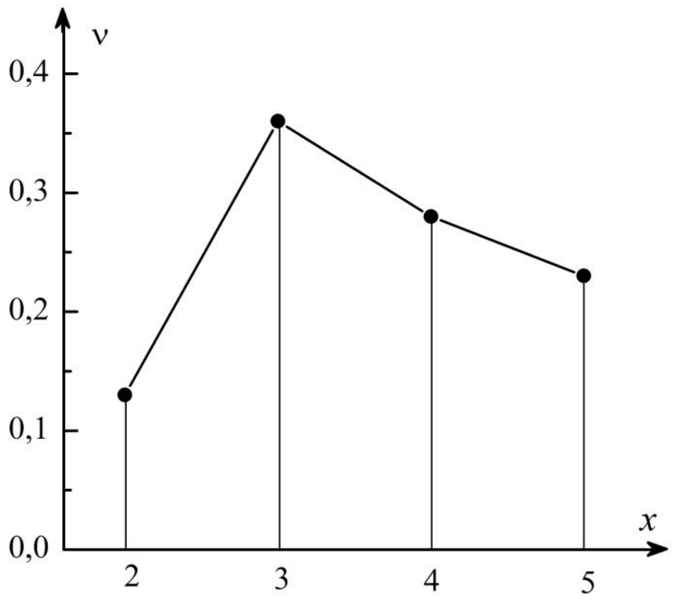


Рис. 4.1. Полигон, соответствующий вариационному ряду, заданному табл. 4.1.

Для интервального вариационного ряда графическое изображение называется гистограммой. Каждому разряду на гистограмме соответствует прямоугольник длина основания которого равна длине разряда, а высота – соответствующей частоте m_i или относительной частоте ν_i. Но обычно относительную частоту разряда делят на его длину. Тогда полная площадь гистограммы будет равна единице. При такой «нормировке» гистограмма является статистическим аналогом графика плотности распределения вероятности.

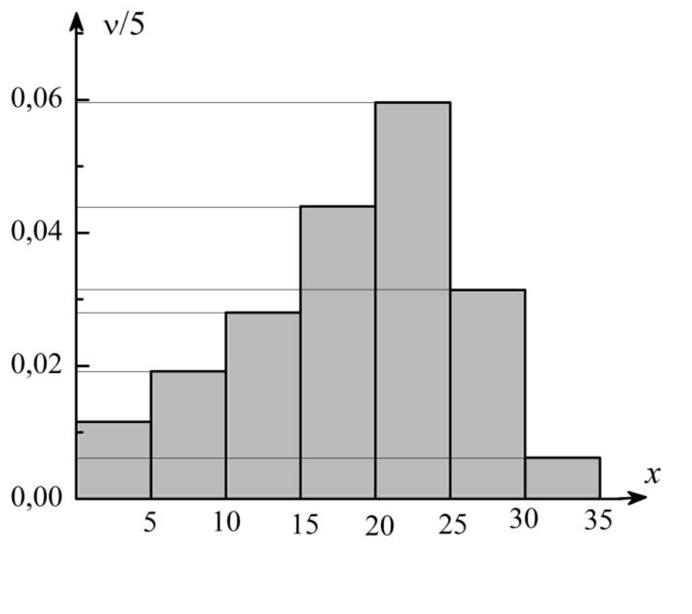


Рис.4.2. Гистограмма, соответствующая вариационному ряду, заданному табл. 4.2.

На рис. 4.2 изображена гистограмма для статистического ряда из таблицы 4.2. По оси ординат отложены относительные частоты, деленные на длину разряда, равную пяти.

Накопленной частотой n_i называется величина, показывающая, сколько имеется значений признака не превосходящих заданного значения x_i. Соответственно, относительной накопленной частотой (накопленной частостью) называется величина $w_i = n_i / n$.

Частоты, относительные частоты, накопленные частоты и частости для вариационного ряда, заданного табл. 4.1 показаны в таблице 4.3, для вариационного ряда из таблицы 4.2 – в таблице 4.4.

Таблица 4.3

x_i	2 («неудовлетворительно»)	3 («удовлетворительно»)	4 («хорошо»)	5 («отлично»)
m_i	13	36	28	23
ν_i	0,13	0,36	0,28	0,23
n_i	13	49	77	100
w_i	0,13	0,49	0,77	1

Таблица 4.4

x_i	0–5	5–10	10–15	15–20	20–25	25–30	30–35
m_i	58	96	140	220	298	157	31
ν_i	0,058	0,096	0,140	0,220	0,298	0,157	0,031
n_i	58	154	294	514	812	969	1000
w_i	0,058	0,154	0,294	0,514	0,812	0,969	1

График, по оси абсцисс которого отложены значения признака, а по оси ординат – накопленные частоты n_i или накопленные относительные частоты, называется кумулятивной кривой или кумулятой.

Если кумулятивная кривая строится для интервального вариационного ряда, то абсцисса первой точки равна началу первого интервала, соответствующая ордината равна нулю, абсцисса второй точки равна началу второго интервала, а ордината равна первой накопленной (относительной) частоте и т.д.

Кумулятивная кривая, построенная по данным таблицы 4.3, показана на рис. 4.3.

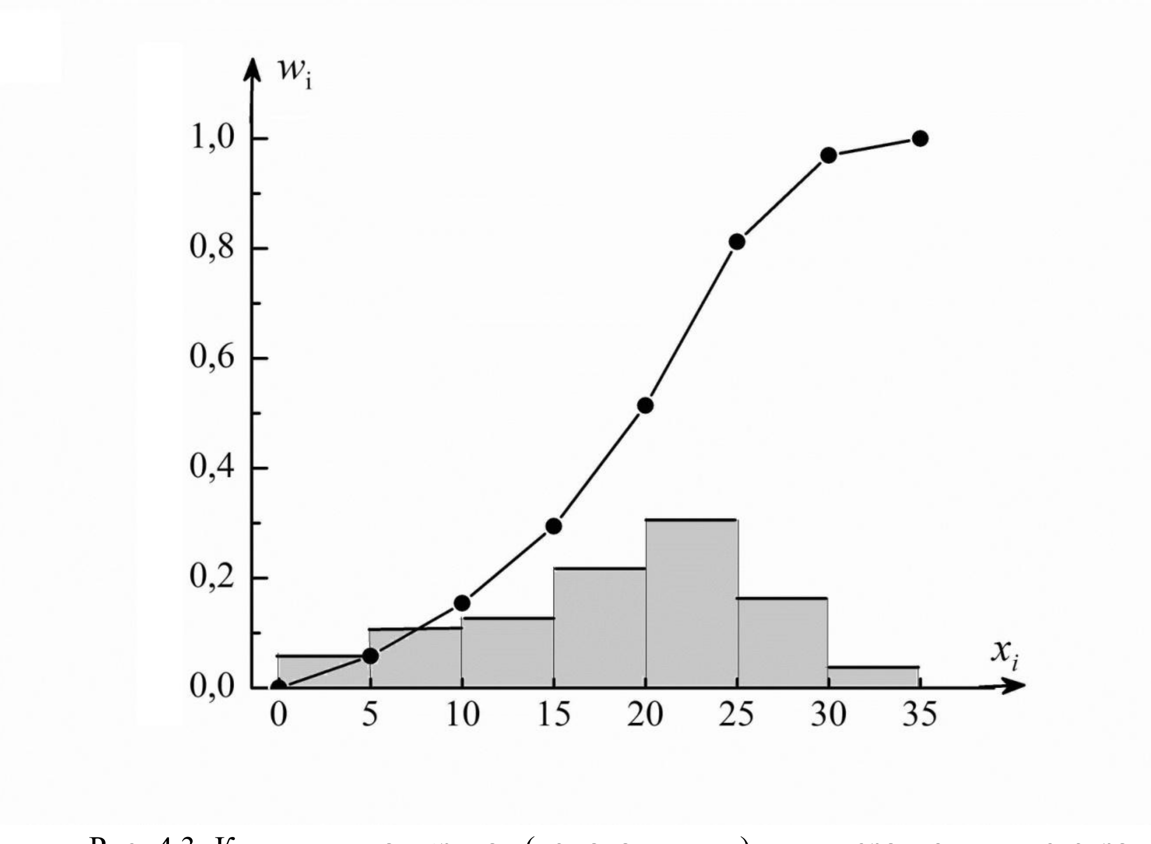


Рис. 4.3. Кумулятивная кривая (ломаная линия) и, для сравнения, гистограмма для статического ряда из табл. 3.4

Статистической или эмпирической функцией распределения называется функция, определяемая следующим образом

$$\widetilde{F}(x) = \begin{cases} 0, & x < x_1 \\ w_k, & x_k \le x < x_{k+1}, \ k = 1,2,\dots,n-1. \\ 1, & x \ge x_n \end{cases} \qquad (4.2)$$

Здесь во второй строке правой части стоит накопленная относительная частота w_k:

$$w_k = \sum_{i=1}^{k} \nu_i, \quad k = 1,2,\ldots,n-1.$$

Значения x_1, x_2, ..., x_n соответствуют значениям признака для дискретного статистического ряда и границам разрядов для интервального ряда.

График статистической функции распределения всегда имеет ступенчатый вид. Но если его «сгладить», соединив прямыми линиями какие либо точки каждой ступеньки, то получим график близкий по виду и смыслу к кумулятивной кривой.

4.2. Оценка параметров генеральной совокупности по ограниченному числу опытов.

Пусть есть некоторая характеристика или параметр a случайной величины, которую мы оцениваем по выборке и получаем таким образом оценку $\tilde{a}$. Следует иметь в виду, что $\tilde{a}$ сама является случайной величиной. Предъявим к $\tilde{a}$ следующие требования.

1. При увеличении числа опытов n $(n \to \infty)$ $\tilde{a}$ должна сходиться по вероятности к истинному значению a: $\tilde{a} \xrightarrow{P} a$. Оценка, обладающая таким свойством, называется *состоятельной*.

2. Математическое ожидание случайной величины $\tilde{a}$ должно быть равно a: $M[\tilde{a}] = a$. Оценка, обладающая таким свойством, называется *несмещенной*.

Прежде всего, нас интересует оценка математического ожидания m_x случайной величины *X*. ***Состоятельной и несмещенной оценкой математического ожидания является среднее арифметическое***, которое далее мы будем обозначать $\tilde{m}_x$:

$$\tilde{m}_x = \frac{1}{n}\sum_{i=1}^{n} x_i . \tag{4.3}$$

Действительно, непосредственно из закона больших чисел следует, что при $n \to \infty$ $\tilde{m}_x \xrightarrow{P} m_x$.

Найдем теперь математическое ожидание среднего арифметического.

$$M[\tilde{m}_x] = M\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] = \frac{1}{n}\sum_{i=1}^{n} M[x_i] = \frac{1}{n}\sum_{i=1}^{n} m_x = \frac{1}{n}\cdot n\cdot m_x = m_x. \tag{4.4}$$

Здесь мы воспользовались двумя свойствами математического ожидания: постоянный множитель можно выносить за знак математического ожидания и математического ожидание суммы равно сумме математических ожиданий.

Обратимся теперь к дисперсии. Дисперсия – это тоже среднее значение (математическое ожидание), но, конечно, не самой случайной величины *X*, а величины $(X - m_x)^2$: $D_x = D[X] = M\left[(X - m_x)^2\right]$. Поэтому представляется естественным за оценку дисперсии принять среднее арифметическое квадратов отклонений выборочных значений от среднего арифметического:

$$\tilde{D}_x = \frac{1}{n}\sum_{i=1}^{n}(x_i - \tilde{m}_x)^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \tilde{m}_x^2 . \tag{4.5}$$

Таким образом, данная оценка дисперсии равна среднему арифметическому квадратов выборочных значений минус квадрат среднего арифметического. Из закона больших чисел автоматически следует, что такая оценка является состоятельной.

Выясним, является ли такая оценка несмещенной. Можно показать, что

$$M[\tilde{D}_x] = \frac{n-1}{n} D_x.$$

Поэтому за состоятельную и несмещенную оценку дисперсии принимается величина

$$\tilde{D}_x = \frac{n}{n-1} \cdot \frac{1}{n} \sum_{i=1}^{n} (x_i - \tilde{m}_x)^2,$$

т.е.

$$\tilde{D}_x = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \tilde{m}_x)^2 = \frac{1}{n-1} \sum_{i=1}^{n} x_i^2 - \frac{n}{n-1} \tilde{m}_x^2. \qquad (4.6)$$

При больших объемах выборки разница между оценками (4.5), (4.6) невелика и практически не важно какую из них использовать. Но при небольших n следует пользоваться выражением (4.6).

В статистике среднее арифметическое (4.3) и выражение (4.5), в предположении, что они вычислены для генеральной совокупности, называются генеральной средней и генеральной дисперсией. Из сказанного выше следует, что они полностью аналогичны математическому ожиданию и дисперсии в теории вероятностей. Оценки, сделанные по выборке, называются выборочным средним и выборочной дисперсией. Далее, для генеральных средней и дисперсии сохраним обозначения m_x, D_x, для выборочных величин мы будем использовать обозначения, снабженные тильдой: $\tilde{m}_x, \tilde{D}_x$.

Оценки параметров распределений, которые выражаются одним числом, называются точечными оценками. Таким образом, выражения (4.3), (4.5) или (4.6) являются точечными оценками математического ожидания и дисперсии.

Одна из основных задач математической статистики заключается в определении того, насколько точечные оценки параметров распределений близки к истинным, под которыми понимаются соответствующие параметры для генеральной совокупности.

Рассмотрим данную проблему для математического ожидания. В статистике решаются два вида задач, связанных с оценкой математического ожидания:

- какова вероятность того, что математическое ожидание отклонится от среднего арифметического на величину не более, чем заданная величина ε;
- каков интервал $(\tilde{m}_x - \varepsilon, \tilde{m}_x + \varepsilon)$ в который математическое ожидание попадает с заданной вероятностью, которую мы будем обозначать как γ.

Интервал $(\tilde{m}_x - \varepsilon, \tilde{m}_x + \varepsilon)$ называется доверительным интервалом для математического ожидания, вероятность γ – доверительной вероятностью.

Решение поставленных задач основано на Центральной Предельной Теореме. Действительно, среднее арифметическое вычисляется как сумма и, следовательно, в предположении, что измерения, в результате которых получены выборочные значения, являются независимыми, закон распределения среднего арифметического близок к нормальному при достаточно большом числе слагаемых. Распределение признака генеральной совокупности здесь не важно.

Составим случайную величину

$$Y_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{x_i - m_x}{\sigma_x},$$

где $\sigma_x = \sqrt{D_x}$ – среднее квадратическое отклонение.

Согласно ЦПТ распределение данной СВ близко к нормальному с нулевым математическим ожиданием и единичной дисперсией. Поэтому

$$P(-\eta < Y_n < \eta) = 2\tilde{\Phi}(\eta).$$

$$Y_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{x_i - m_x}{\sigma_x} = \frac{1}{\sqrt{n}} \cdot \frac{\sum_{i=1}^{n} x_i - n m_x}{\sigma_x} = \frac{n}{\sqrt{n}} \cdot \frac{\frac{1}{n}\sum_{i=1}^{n} x_i - m_x}{\sigma_x} = \frac{\tilde{m}_x - m_x}{\sigma_x / \sqrt{n}}.$$

Следовательно,

$$P\left(-\eta < \frac{\tilde{m}_x - m_x}{\sigma_x / \sqrt{n}} < \eta\right) = 2\Phi(\eta), \text{ где } \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-t^2/2} dt$$

Пусть $\varepsilon \equiv \eta \cdot \frac{\sigma_x}{\sqrt{n}}$, тогда

$$P(-\varepsilon < \tilde{m}_x - m_x < \varepsilon) = 2\Phi\left(\frac{\varepsilon}{\sigma_x / \sqrt{n}}\right).$$

Как правило, истинное среднее квадратическое отклонение априори неизвестно. Поэтому последнее, что надо сделать, заменить истинную дисперсию $D_x = \sigma_x^2$ (которую мы не знаем) ее оценкой $\tilde{D}_x$. Окончательно получим формулу, которая решает поставленные выше задачи

$$P(-\varepsilon < \tilde{m}_x - m_x < \varepsilon) = 2\Phi\left(\frac{\varepsilon}{\sqrt{\tilde{D}_x / n}}\right) \text{ или } P(|m_x - \tilde{m}_x| < \varepsilon) = 2\Phi\left(\frac{\varepsilon}{\sqrt{\tilde{D}_x / n}}\right). \qquad (4.7)$$

Если задана доверительная вероятность γ, то по таблице функций Лапласа определяют соответствующее значение аргумента t_γ:

$$\Phi(t_\gamma) = \gamma / 2.$$

После чего получаем искомое значение ε:

$$\varepsilon = t_\gamma \cdot \sqrt{\frac{\tilde{D}_x}{n}}. \qquad (4.8)$$

Примеры.

1. При обработке $n = 100$ независимых опытов получены следующие оценки генеральной средней и генеральной дисперсии: $\tilde{m}_x = 4{,}52$, $\tilde{D}_x = 2{,}35$.

Найти

а) интервал, в которой генеральное среднее попадает с вероятностью γ=0,99;

б) вероятность, что генеральное среднее отклонится от выборочного среднего на величину не более чем $\varepsilon = 0{,}3$.

Решение

а) Используем формулу (4.7):

$$2\Phi\left(\frac{\varepsilon}{\sqrt{\tilde{D}_x/n}}\right)=\gamma,\ \Phi\left(\frac{\varepsilon}{\sqrt{2{,}35/100}}\right)=0.495.$$

По таблице функции Лапласа $\Phi(x)$ определяем значение аргумента, соответствующее значению функции 0,475, Это значение примерно равно 2,58:

$$\frac{\varepsilon}{\sqrt{2.35/100}}\approx 2{,}58,\ \varepsilon=2{,}58\cdot\sqrt{\frac{2{,}35}{100}}\approx 0{,}396.$$

Для доверительного интервала получаем $(4{,}52-0{,}396;\ 4{,}52+0{,}396)$, $(4{,}12;\ 4{,}92)$.

б) Здесь по таблице функций Лапласа нам надо определить ее значение при аргументе равном $\frac{0{,}3}{\sqrt{2{,}35/100}}\approx 1{,}96$: $\Phi(1{,}96)=0{,}475$. Для искомой вероятности получим:

$$P(|m_x - \tilde{m}_x|<\varepsilon)=2\cdot 0{,}475=0{,}95.$$

2. Для исследования доходов населения города отобрано 1000 человек. Получено следующее распределение жителей по месячному доходу (тыс.руб)

Таблица 4.5

Доход x_i тыс.руб.	Менее 5	5–10	10–15	15–20	20–25	Свыше 25
	2,5	7,5	12,5	17,5	22,5	27,5
Число жителей m_i	58	96	239	328	147	132

а) Найти вероятность того, что средний доход отличается от выборочного среднего не более, чем на 450 руб=0,45 тыс.руб.

б) Определить границы, в которых с вероятностью γ=0,99 заключен средний доход.

Прежде всего, видим, что в первой строке таблицы находятся не сами значения дохода, а интервалы в которых эти значения находятся. Для дальнейших вычислений необходимо выбрать так называемого «представителя» каждого интервала. В качестве такового обычно выбирают середины интервалов. Середины интервалов записаны во второй строке таблицы. Для последнего промежутка отсутствует правая граница, но так как значения представителей следуют с шагом 5, приписываем последнему представителю значение 27,5.

Вычисляем среднее арифметическое

$$\tilde{m}_x=\frac{1}{1000}\sum_{i=1}^{1000}x_i=\frac{1}{1000}\sum_{k=1}^{6}x_k m_k.$$

Так как отношение $m_k/1000$ это относительная частота ν_k, формулу для среднего арифметического можно переписать в виде

$$\tilde{m}_x=\sum_{k=1}^{6}x_k\nu_k.$$

Приходим к общему важному выводу: формула для оценки математического ожидания (т.е. для среднего арифметического) может быть получена из формулы для математического ожидания, но с заменой вероятности на относительную частоту.

Вообще, можно доказать (попытайтесь сделать это самостоятельно) следующее важное утверждение: ***относительная частота есть состоятельная и несмещенная оценка вероятности***.

Возвращаемся к вычислениям.

$$\tilde{m}_x = \frac{1}{1000}(2{,}5\cdot 58 + 7{,}5\cdot 96 + 12{,}5\cdot 239 + 17{,}5\cdot 328 + 22{,}5\cdot 147 + 27{,}5\cdot 132) = 16{,}53.$$

Дисперсию будем вычислять смещенную, т.е используем формулу (4.5), так как объем выборки здесь достаточно велик.

$$\tilde{D}_x = \frac{1}{1000}\left(2{,}5^2\cdot 58 + 7{,}5^2\cdot 96 + 12{,}5^2\cdot 239 + 17{,}5^2\cdot 328 + 22{,}5^2\cdot 147 + 27{,}5^2\cdot 132\right) -$$

а)

$$-16{,}53^2 \approx 44{,}56.$$

Используем формулу (4.7)

$$P(|m_x - \tilde{m}_x| < 0{,}45) = 2\Phi\left(\frac{0{,}45}{\sqrt{44{,}56/1000}}\right) = 2\cdot 0{,}48341 \approx 0{,}967$$

б) Теперь нам неизвестна величина ε, но известна вероятность γ=0,99:

$$P(|m_x - \tilde{m}_x| < \varepsilon) = \gamma$$

или $2\Phi\left(\frac{\varepsilon}{\sqrt{44{,}56/1000}}\right) = 0{,}99$, $\Phi\left(\frac{\varepsilon}{\sqrt{44{,}56/1000}}\right) = 0{,}495$.

По таблице функции Лапласа определяем значение аргумента, соответствующее значению функции 0,495:

$$\frac{\varepsilon}{\sqrt{44{,}56/1000}} \approx 2{,}58,\ \varepsilon \approx 0{,}55.$$

И получаем искомый интервал $(16{,}53 - 0{,}55; 16{,}53 + 0{,}55) = (15{,}98; 17{,}08)$.

Интуитивно ясно, что нормальное распределение для среднего арифметического можно использовать с хорошей точностью в том случае, когда объем выборки достаточно велик. Если объем выборки невелик (условно, n<50) используют распределение Стьюдента. А именно, доказано, что если случайная величина X (в статистике говорят также признак X генеральной совокупности) распределена по нормальному закону, то величина

$$T_n = \frac{\tilde{m}_x - m_x}{\sqrt{\tilde{D}_x / n}} \qquad (4.9)$$

подчиняется распределению Стьюдента с $f = n - 1$ степенями свободы. В (4.9) $\tilde{D}_x$ несмещенная оценка дисперсии, вычисляемая по формуле (4.6). Плотность распределения Стьюдента $S_f(x)$ является четной функцией. Поэтому вероятность β попадания случайной величины T_n в интервал $(-t_\beta, t_\beta)$ может быть определена как

$$\beta = 2\int_0^{t_\beta} S_f(x)dx.$$

Обычно для распределения Стьюдента в литературе по теории вероятностей и математической статистике приводят таблицу, в которой даются значения t_β в зависимости от вероятности β. Определив по такой таблице t_β, для величины ε получим

$$\varepsilon = t_\beta \sqrt{\frac{\tilde{D}_x}{n}}\,. \qquad (4.10)$$

Таблица распределения Стьюдента приведена в приложении 2.

Пример.

При обработке $n = 20$ независимых опытов получены оценка генеральной средней и несмещенная оценка дисперсии: $\tilde{m}_x = 4{,}52$, $\tilde{D}_x = 2{,}35$.

Найти интервал, в которой генеральное среднее попадает с вероятностью 0.99, а) используя распределение Лапласа, б) распределение Стьюдента.

Решение

а) По таблице функций Лапласа находим значение аргумента t_γ, соответствующее значению функции 0,99/2=0,495: $t_\gamma = 2{,}58$. Следовательно,

$$\varepsilon = 2{,}58 \cdot \sqrt{\frac{2{,}35}{20}} \approx 0{,}884.$$

Искомый интервал $(3{,}64;\, 5{,}40)$.

б) По таблице распределения Стьюдента с f=19 степенями свободы находим значение t_β: $t_\beta = 2{,}8609$. Теперь определяем величину ε:

$$\varepsilon = 2{,}8609 \cdot \sqrt{\frac{2{,}35}{20}} \approx 0{,}981.$$

Искомый интервал $(3{,}54;\, 5{,}50)$.

Подведем итог. Доверительный интервал для математического ожидания признака генеральной совокупности определяется величиной ε:

$$\varepsilon = t\sqrt{\frac{\tilde{D}_x}{n}}\,, \qquad (4.11)$$

где для достаточно больших выборок коэффициент t определяется с помощью таблицы функции Лапласа, а для небольших – по таблице распределения Стьюдента с $f = n - 1$ степенями свободы.

Рассмотрим задачу оценки по выборке вероятности p появления (иначе относительной доли) определенного значения признака генеральной совокупности. В качестве выражения для генеральной дисперсии следует использовать дисперсию для распределения Бернулли: $D_p = p(1-p)$. Так как относительная частота является состоятельной и несмещенной оценки вероятности, выборочная дисперсия определяется как

$$\tilde{D}_p = \nu(1-\nu)\,. \qquad (4.12)$$

Пример

По проведении социологического исследования было опрошено 1200 респондентов. Опрос показал, что рейтинг президента составил 70%. Найти границы в которых с вероятностью 0,95 заключен действительный рейтинг президента.

Решение

Под действительным рейтингом здесь подразумевается рейтинг, полученный по генеральной совокупности, т.е. при опросе всех потенциальных избирателей. Границы определяются величиной ε, которую определяем по формуле (4.11). Величину t определяем по таблице функции Лапласа: t=1.96. Для оценки дисперсии используем выражение (4.12) с $\nu = 0{,}7$. Таким образом

$$\varepsilon = 1{,}96\sqrt{\frac{0{,}7 \cdot 0{,}3}{1200}} \approx 0{,}026.$$

Границы равны (67,4%; 72,6%).

4.3. Проверка статистических гипотез

Статистической гипотезой называется любое предположение о виде неизвестного закона распределения или значениях параметров распределения. Проверяемую гипотезу будем обозначать H_0. Наряду с H_0 обычно выдвигают конкурирующую гипотезу H_1.

Пусть имеется выборка значений некоторого признака генеральной совокупности: $x_1, x_2, \ldots, x_n$. По выборке определяем некоторый параметр q: $q = \varphi(x_1, x_2, \ldots, x_n)$. В качестве q может выступать выборочное среднее, выборочная дисперсия, другие моменты случайной величины. Параметр q называется статистикой. Предполагается, что известен точный или приближенный закон распределения параметра q и, следовательно, мы можем точно или приближенно любую из вероятностей $P(q > q_c)$, $P(q < q_c)$, $P(|q| < q_c)$ где q_c – определенное значение q.

Пусть Q множество значений параметра q: $q \in Q$. Процесс проверки заключается в том, что все множество значений величины q разбивается на два непересекающихся подмножества: $Q = Q_0 \cup Q_1, Q_0 \cap Q_1 = \varnothing$. Если $q \in Q_0$ гипотеза H_0 принимается, если $q \in Q_1$ H_0 отвергается. Множество Q_1 называется критической областью. Так как q – случайная величина, то отнесение ее к определенной области значений носит вероятностный характер.

Пусть гипотеза H_0 верна, но отвергается в силу того, что найденное значение q попало в область Q_1: $q \in Q_1$. Такая ошибка называется ошибкой первого рода. Вероятность такой ошибки будем обозначать буквой α, вероятность α называется уровнем значимости. Величина 1–α является доверительной вероятностью (это вероятность того, что мы доверяем гипотезе H_0).

Пусть гипотеза H_0 неверна, но принимается в силу того, что найденное значение q попало в область Q_0: $q \in Q_0$. Такая ошибка называется ошибкой второго рода, ее вероятность будем обозначать буквой β.

Подчеркнем, главной проблемой при проверке статистических гипотез является выбор точного или приближенного закона распределения статистики q.

Проверка гипотез об вероятных значениях параметров распределений.

Между определением интервальных оценок и проверкой гипотез об их вероятных значениях существует тесная связь. Фактически, это два альтернативных подхода к решению одной и то же задачи. Рассмотрим задачу об определении доверительного интервала для математического ожидания. Пусть мы предполагаем, что математическое ожидание равно определенному значению m_x (гипотеза H_0). Пусть по выборке получено значение $\tilde{m}_x$ как среднее арифметическое (4.3). Зададимся уровнем значимости, например, $\alpha = 0.05$ (доверительная вероятность 0.95). Областью Q_0 является интервал $(m_x - \varepsilon, m_x + \varepsilon)$, где величина ε определяется из (4.8). Критическая область Q_1 является двусторонней: $(-\infty, m_x) \cup (m_x, +\infty)$. Области Q_0, Q_1 и соответствующие вероятности показаны на рис. 3.4.

Примем ли мы гипотезу H_0 или отвергнем ее зависит от того в какую область попадет значение $\widetilde{m}_x$.

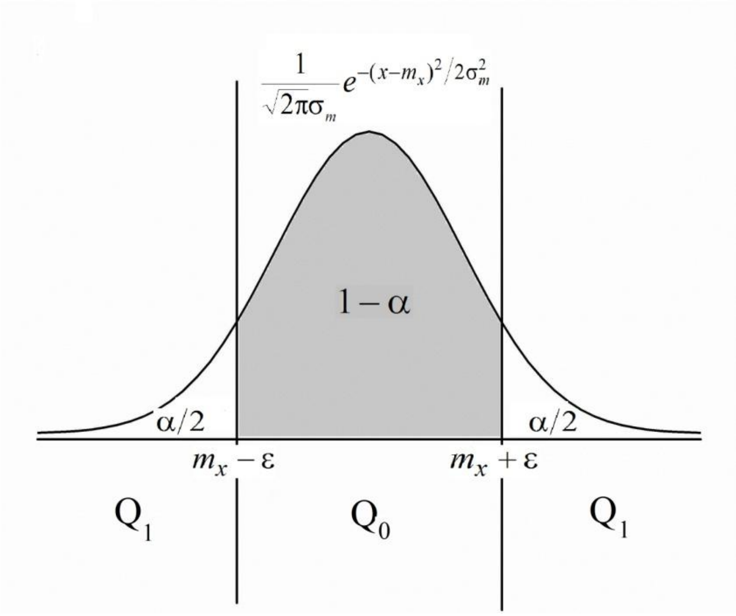


Рис. 4.4.

Примеры

1. На станке в автоматическом режиме изготавливаются шарики для подшипников. Средний диаметр шарика должен составлять 35мм. Однако для выборки из 80 подшипников он составил 35,1мм. При среднем квадратическом отклонении 0,15 мм и при 5% уровне значимости (доверительная вероятность равна 0,95) проверить гипотезу о том, что станок не требует подналадки.

Даже при работе на тщательно налаженном станке размер шарика будет отклоняться от проектного в результате сугубо случайных (или, как говорят, статистических) факторов. Кроме того, сам процесс измерения диаметра готового изделия будет давать некоторую погрешность. Фактически, мы пытаемся выяснить, является ли отклонение диаметра от проектного только статистическим фактором или же имеет место систематическая ошибка.

Гипотеза H_0: $m_x = 35$ мм. Конкурирующая гипотеза H_1: $m_x \neq 35$ мм, критическая область Q_1 является двусторонней. Область Q_0 определяется величиной

$$\varepsilon = t_{0.95}\frac{0,15}{\sqrt{80}}.$$

Коэффициент $t_{0.95}$ найдем по таблице функций Лапласа: $t_{0.95} = 1{,}96$. Таким образом, $\varepsilon \approx 0{,}033$ мм. Область принятия гипотезы Q_0 интервал (34,967; 35,033). Значение 35,1 не принадлежит области Q_0. При данном уровне значимости гипотеза H_0 отвергается. Станок требует подналадки.

2. Средний процент брака готовой продукции, найденный для десяти предприятий отрасли составляет 1,3%, а выборочное среднее квадратическое отклонение (несмещенное) равно 1.1%. Можно ли считать, что средний уровень брака для всех предприятий отрасли не превосходит 1%? Уровень значимости примем равным α=0,05.

Гипотеза H_0: средний уровень брака m_x не превосходит 1%, конкурирующая гипотеза H_1: m_x>1%, область Q_1 является односторонней.

Выборка невелика, поэтому используем распределение Стьюдента. Напомним, что плотность распределения Стьюдента симметрична, а в таблице распределения дается величина t_β, определяющая интервал $(-t_\beta, t_\beta)$ для заданной вероятности β. Так как область Q_1 односороняя, величина β определяется как $\beta = 1 - 2\alpha$ (см.рис 4.5).

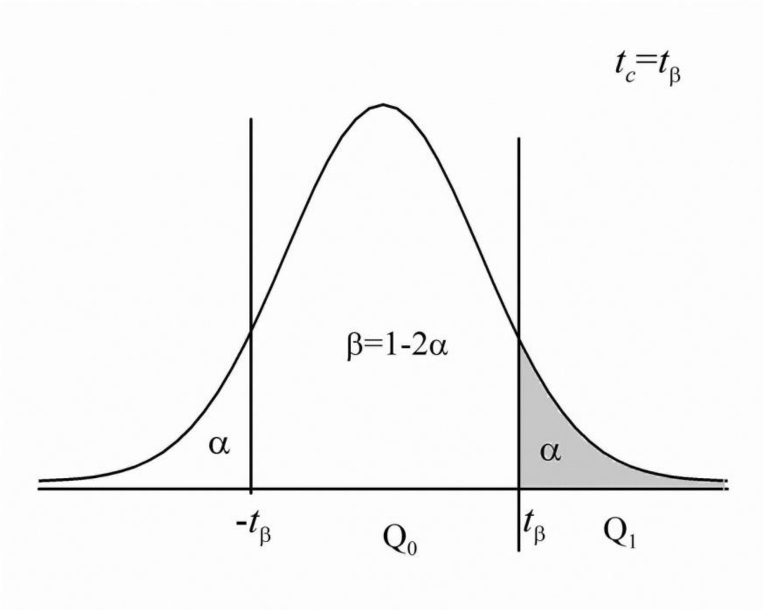


Рис.4.5. Область Q_0 принятия гипотезы H_0, и критическая область Q_1 для конкурирующей гипотезы H_1: $t > t_c$ для нормального распределения и распределения Стьюдента.

Число степеней свободы равно 9. По таблице распределения Стьюдента для β=0,9 определяем t_β=1,8331. Теперь определяем границу критической области ε_c:

$$\varepsilon_c = 1{,}8331 \frac{1{,}1}{\sqrt{10}} \approx 0{,}64\% .$$

В нашем случае отклонение от генерального среднего равно $1{,}3\% - 1\% = 0{,}3\%$, что меньше найденного критического значения. Следовательно, гипотеза H_0 принимается.

Проверка гипотезы о равенстве математических ожиданий (генеральных средних).

Пусть имеются две генеральных совокупности одной и той же природы. Например, населения двух крупных регионов, два поля засеянных пшеницей, детали, изготовленные двумя рабочими на разных станках и т.д. Для генеральных совокупностей мы получаем две выборки некоторого признака: x_1, x_2,... и y_1, y_2,.... Объем первой выборки обозначим как n_1, второй – n_2. Выборочные средние обозначим, соответственно, как $\tilde{m}_x$ и $\tilde{m}_y$. Величины $\tilde{m}_x$, $\tilde{m}_y$ будут различаться, интуитивно ясно, что вероятность того, что они совпадут, крайне мала. Но тогда возникает вопрос: обусловлено ли разность выборочных средних разницей генеральных средних или она носит сугубо статистический характер. Например, мы нашли средний уровень дохода для жителей двух соседних регионов. Говорит ли разница в уровнях дохода о действительно разных уровнях жизни в этих регионах или же уровень жизни одинаков, а разница выборочных средних имеет случайный характер. Пусть H_0 – гипотеза о равенстве генеральных средних: $m_x = m_y$; H_1 – конкурирующая гипотеза.

Составим случайную величину

$$q=\frac{\tilde{m}_x-\tilde{m}_y-\mathrm{M}\lfloor\tilde{m}_x-\tilde{m}_y\rfloor}{\sqrt{\mathrm{D}[\tilde{m}_x-\tilde{m}_y]}}.$$

Как уже отмечалось выше, при условии независимости измерений, выборочное среднее имеет распределение близкое к нормальному при достаточно большой выборке.

Если справедлива гипотеза H_0, то

$$\mathrm{M}\lfloor\tilde{m}_x-\tilde{m}_y\rfloor=\mathrm{M}[\tilde{m}_x]-\mathrm{M}\lfloor\tilde{m}_y\rfloor=m_x-m_y=0.$$

Так как дисперсия суммы (разности) независимых случайных величин равна сумме соответствующих дисперсий, то теперь для величины q несложно получить

$$q=\frac{\tilde{m}_x-\tilde{m}_y}{\sqrt{\dfrac{D_x}{n_1}+\dfrac{D_y}{n_2}}}, \qquad (4.13)$$

причем распределение величины q близко к нормальному с нулевым математическим ожиданием и единичной дисперсией т.е. функцией распределения q является функция

$$\Phi^*(q)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{q}e^{-t^2/2}dt=\frac{1}{2}+\Phi(q).$$

Как правило, дисперсии D_x, D_y неизвестны, поэтому вместо них берут выборочные значения.

Зададимся определенным уровнем значимости α. Если конкурирующая гипотеза H_1 определяется условием $m_x\neq m_y$, то критическая область будет двусторонней, она определяется условием $|q|>q_c$. Значение q_c определяется из уравнения (см. рис.4.4):

$$2\Phi(q_c)=1-\alpha \qquad (4.14)$$

Если найденное по формуле (4.13) значение удовлетворяет неравенству $|q|<q_c$ – гипотеза H_0 принимается, в противном случае отвергается.

Примеры

1. Выборочное среднее значение результирующей оценки по дисциплине «математика» для ВУЗа равно 3,90 , а для его филиала 3,82. Объемы выборок равны 110 и 80, выборочные несмещенные дисперсии 1,05 и 1,06 соответственно. Для уровня значимости α=10% проверить гипотезу об одинаковой средней оценке по математике в ВУЗе и филиале.

Решение.

Конкурирующая гипотеза H_1: $m_x\neq m_y$, критическая область является двусторонней.

Значение статистики (4.13) равно

$$q=\frac{3{,}90-3{,}82}{\sqrt{\dfrac{1{,}05}{110}+\dfrac{1{,}06}{80}}}\approx 0{,}53.$$

Значение q_c определяем из условия $2\Phi(q_c)=0.9$ или $\Phi(q_c)=0{,}45$: q_c=1,64. Так как $|q|<q_c$ гипотеза об одинаковой средней оценке принимается.

Пусть конкурирующая гипотеза H_1: $m_x>m_y$. Тогда критическая область будет односторонней. Значение q_c в этом случае определяется из условия

$$\Phi^*(q_c)=1-\alpha \text{ или } \Phi(q_c)=\frac{1}{2}-\alpha.$$

2. Выборочное среднее значение результирующей оценки по профилирующему предмету для ВУЗа равно 3,94 , а для его филиалов 3,69. Объемы выборок равны 100 и 200, выборочные несмещенные дисперсии 1,05 и 1,00 соответственно. Можно ли считать при уровне значимости α=5%, что успеваемость по профилирующему предмету выше в ВУЗе, чем в его филиалах?

Решение.

Конкурирующая гипотеза H_1: $m_x > m_y$ (средняя оценка выше в ВУЗе)

Значение статистики (4.13) равно

$$q = \frac{3{,}94 - 3{,}69}{\sqrt{\dfrac{1{,}05}{100} + \dfrac{1{,}00}{200}}} \approx 2{,}01$$

Критическое значение q_c определяем из условия $\Phi(q_c) = 0{,}45$ ($0{,}5 - 0{,}05 = 0{,}45$): q_c=1,64. Так как $q > q_c$ принимается конкурирующая гипотеза, то есть при данном уровне значимости можно считать, что успеваемость в ВУЗе выше, чем в его филиалах.

Для конкурирующей гипотезы H_1: $m_x < m_y$ все построения совершенно аналогичны.

Проверка гипотезы о равенстве генеральных относительных частот (долей).

Пусть имеется две генеральных совокупности, для которых получены выборки. Проверяем гипотезу H_0 о равенстве относительных генеральных частот (долей) некоторого признака для данных совокупностей: $p_1 = p_2 = p$. Величину p_1 (p_2) можно рассматривать как вероятность получить в данном измерении определенное значение признака.

Относительную частоту признака, определенную по выборке для первой совокупности обозначим как ν_1, по второй – ν_2. Генеральная дисперсия – это, конечно, дисперсия распределения Бернулли. Поэтому в качестве критерия q берем величину

$$q = \frac{\nu_1 - \nu_2}{\sqrt{p(1-p)\cdot\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}},$$

где n_1, n_2 – объемы первой и второй выборок соответственно. Но значение генеральной доли p неизвестно. Обычно, для него используют оценку

$$\tilde{p} = \frac{m_1 + m_2}{n_1 + n_2},$$

где m_1 – частота появления признака в первой выборке, m_2 – во второй. Далее, все вычисления точно следуют общей схеме проверки гипотезы о равенстве генеральных средних.

Проверка гипотезы о равенстве генеральных дисперсий.

Сравнение дисперсий является важной задачей математической статистики.

Поясним это на примерах.

Пусть имеется две группы студентов. Средняя оценка по некоторому предмету, вычисленная для обеих групп, одинакова. Но дисперсия оценки для первой группы выше. Интуитивно понятно, что в первой группе число плохо успевающих студентов выше.

Пусть имеется две группы предприятий некоторой отрасли. Средняя доходность по акциям обеих групп одинакова. Но дисперсия доходности выше для первой группы. Это значит, что выше и риск для акций первой группы.

Из общих рассуждений, приведенных выше, следует, что для того чтобы сделать заключение о равенстве (или неравенстве) дисперсий D_x, D_y двух генеральных совокупностей, нам нужно найти параметр, который связывал бы значения выборочных дисперсий и закон распределения которого был бы известен.

Пусть имеет место гипотеза H_0: $D_x = D_y$. Пусть $\tilde{D}_x, \tilde{D}_y$ – выборочные *несмещенные* дисперсии значений некоторого признака для первой и второй совокупностей. В предположении, что изучаемый признак X генеральных совокупностей распределен нормально, величина

$$F = \frac{\tilde{D}_x}{\tilde{D}_y} \tag{4.15}$$

имеет распределение, вид которого известен. Это распределение называется F-распределением или распределением Фишера–Снедекора. Оно зависит от числа степеней свободы для первой и второй выборок: $f_1 = n_1 - 1$, $f_2 = n_2 - 1$. В руководствах по теории вероятностей и математической статистике в таблицах F-распределения приводятся аргументы функции $\psi(F_\alpha) = P(F > F_\alpha)$ для нескольких значений α этой функции (α – уровень значимости). Чтобы таблицы не занимали слишком много места (распределение зависит от двух параметров f_1, f_2), часто приводятся только несколько уровней значимости: (например, 0.05, 0.01). *Важно*: *таблицы составлены таким образом, что при вычислении значения F в числителе (4.15) следует брать большую из дисперсий.*

В отличие от нормального распределения и распределения Стьюдента плотность распределения величины *F* несимметрична относительно математического ожидания. На рис. 4.6 показаны критические области для различных конкурирующих гипотез и соответствующие значения величины *F*.

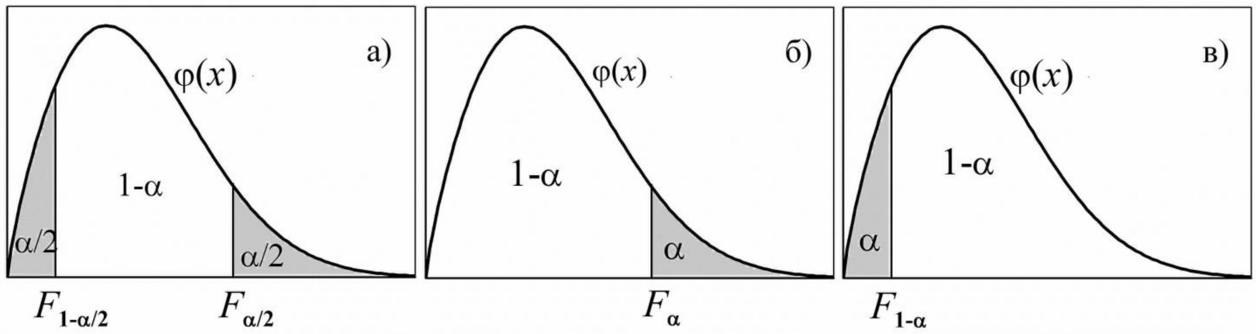


Рис. 4.6. Критические области для F-распределения (закрашены серым цветом), на горизонтальной оси отмечены соответствующие значения величины *F*.

а) двусторонняя область, конкурирующая гипотеза H_1: $D_x \neq D_y$;

б) правосторонняя область, конкурирующая гипотеза H_1: $D_x > D_y$;

в) левосторонняя область, конкурирующая гипотеза H_1: $D_x < D_y$.

Для вычисления значений аргумента F-распределения в зависимости от уровня значимости α и числа степеней свободы f_1, f_2 рекомендуем пользоваться функцией FРАСПОБР(α; f_1; f_2) MS Excel.

Пример

В ВУЗе получены данные об успеваемости студентов по дисциплине «математика» для первого и второго курсов. Выборочная дисперсия оценки по дисциплине для студентов первого курса составила 1,30 при величине выборки n_1=25, для студентов второго курса выборочная дисперсия 1,52 при выборке n_2=20. При уровне значимости α=5% проверить гипотезу о равенстве дисперсий.

Решение.

Как конкурирующую гипотезу H_1 возьмем гипотезу, что дисперсия оценки для второго курса выше, чем для первого. Находим значение статистики F

$$F = \frac{1{,}52}{1{,}30} \approx 1{,}17.$$

Для числителя число степеней свободы равно f_1=19, для знаменателя f_2=24. Критическое значение F для данного уровня значимости равно F_c≈2,04. Так как $F < F_c$ гипотеза H_0 о равенстве дисперсий не отвергается.

Вычисления при конкурирующей гипотезе H_1: $D_x \neq D_y$ (двусторонняя критическая область) осложняются тем обстоятельством, что, как отмечалось выше, плотность распределения несимметрична. Но можно воспользоваться следующим свойством F-распределения:

$$F_{1-\alpha/2}(f_2,\, f_1) = \frac{1}{F_{\alpha/2}(f_1,\, f_2)}. \qquad (4.16)$$

Проверка гипотезы о распределении признака X генеральной совокупности.

Здесь рассматривается задача об определении из опытных данных, т.е. по выборке $x_1, x_2, \ldots, x_n$ значений некоторого признака X, закона распределения этого признака.

Более конкретно, при заданном уровне значимости нам нужно проверить гипотезу H_0 о том, что данный признак имеет определенный закон распределения.

Пусть ν_i, i=1,2,..., k, относительные частоты наблюденных значений, p_i – вероятности этих значений, вычисленные в соответствии с предполагаемым законом распределения.

Степень отклонения эмпирических частот от вероятностей можно охарактеризовать одним числом q, которое, таким образом, является мерой расхождения эмпирического и теоретического распределений:

$$q = \sum_{i=1}^{k} c_i (\nu_i - p_i)^2, \qquad (4.17)$$

где весовые коэффициенты c_i – неизвестные пока числа, которые будут определены ниже. Если бы мы знали закон распределения величины q, задача была бы решена.

К.Пирсон* показал, что если выбрать $c_i = n/p_i$, то с увеличением n закон распределения величины q все меньше и меньше зависит от распределения признака X, приближаясь ко вполне определенному закону, который называется распределением χ^2 («хи квадрат»). При данном выборе коэффициентов c_i мера расхождения также обозначается как χ^2:

$$\chi^2 = n \sum_{i=1}^{k} \frac{(\nu_i - p_i)^2}{p_i}. \qquad (4.18а)$$

Так как $\nu_i = m_i / n$, данное выражение можно записать в виде

* Карл Пирсон (1857–1936) – английский математик, широко использовал методы математической статистики в биологических исследованиях. Рассматриваемый здесь критерий близости распределений часто называют критерием Пирсона.

$$\chi^2 = \sum_{i=1}^{k} \frac{(m_i - np_i)^2}{np_i}. \qquad (4.18б)$$

Распределение χ^2 зависит от числа степеней свободы, которое определяется как $f = k - b$, где b – число связей (дополнительных условий), накладываемых на частоты. Существует, по крайней мере, одно условие, которому должны удовлетворять частоты ν_i:

$$\sum_{i=1}^{k} \nu_i = 1, \qquad (4.19)$$

поэтому минимальное значение для числа b – единица.

В таблице распределения χ^2 приводятся аргументы функции $\psi(\chi^2_\alpha) = P(\chi^2 > \chi^2_\alpha)$ для различных значений уровня значимости α. Таблица распределения дана в приложении 3.

Чем меньше найденное значение χ^2 – тем лучше, поэтому альтернативной гипотезой здесь всегда является гипотеза H_1: $\chi^2 > \chi^2_c$, где χ^2_c – критическое значение критерия, найденное по таблице для данного α.

Любая функция распределения всегда зависит от некоторых параметров. Например, для равномерного распределения параметрами являются числа a, b – границы отрезка, которому принадлежит случайная величина, для распределения Пуассона – параметр a, который одновременно является математическим ожиданием и дисперсией, для нормального распределения – параметры m_x, σ_x, которые являются математическим ожиданием и средним квадратическим отклонением. Эти параметры определяются непосредственно по наблюденным значениям признака X. Каждый найденный параметр является дополнительной связью.

Примеры.

1. Имеются следующие статистические данные по количеству обращений в пункты обслуживания и ремонта в течении часа

Таблица 4.6

Количество обращений, x_i	0	1	2	3	4	5	6
Количество пунктов, m_i	15	22	30	18	10	3	2

Общее количество пунктов n=100. Для уровня значимости α=0,05 проверим гипотезу о том, что величина X – количество обращений, распределена по закону Пуассона.

Найдем выборочное среднее и выборочную дисперсию.

$$\tilde{m}_x = \frac{1}{n}\sum_{i=1}^{7} x_i m_i = \frac{1}{100}(0\cdot 15 + 1\cdot 22 + 2\cdot 30 + 3\cdot 18 + 4\cdot 10 + 5\cdot 3 + 6\cdot 2) = 2.03,$$

$$\tilde{D}_x = \frac{1}{n-1}\sum_{i=1}^{7} x_i^2 m_i - \frac{n}{n-1}\tilde{m}_x^2 =$$

$$= \frac{1}{99}(0\cdot 15 + 1\cdot 22 + 4\cdot 30 + 9\cdot 18 + 16\cdot 10 + 25\cdot 3 + 36\cdot 2) - \frac{100}{99}\cdot 2.03^2 \approx 2.01$$

Для распределения Пуассона математическое ожидание и дисперсия равны: $m_x = D_x = a$. Поэтому для параметра a выбираем значение a=2,02.

Значения вероятностей p_i вычисляются в соответствии с распределением Пуассона для значений $m = x_i$

$$p_i = P(X = m) = \frac{a^m}{m!} e^{-a} . \qquad (4.20)$$

В том случае, если при больших или малых значениях признака значение частот малы, объединяют разряды. Это делается с тем, чтобы уменьшить влияние статистических ошибок на качество описания. Здесь мы объединяем два последних разряда, приписывая значению x_6=5 частоту 5.

В таблице 4.7 представлены значения всех необходимых для вычисления величин. Полигоны для эмпирического и теоретического распределений представлены на рис.3.7.

Таблица 4.7

x_i	0	1	2	3	4	5
частота m_i	15	22	30	18	10	5
Вероятность p_i	0,13	0,27	0,27	0,18	0,09	0,04
np_i	13	27	27	18	9	4

По формуле (4.18б) вычисляем χ^2:

$$\chi^2 = \frac{(15-13)^2}{13} + \frac{(22-27)^2}{27} + \frac{(30-27)^2}{27} + \frac{(18-18)^2}{18} + \frac{(10-9)^2}{9} + \frac{(5-4)^2}{4} \approx 1.93 .$$

Так как по наблюденным данным мы определили *один* параметр a, число связей равно b=2, а число степеней свободы $f = 6 - 2 = 4$. По таблице распределения χ^2 определяем критическое значение статистики: $\chi_c^2 = 9,49$. Так как $\chi^2 < \chi_c^2$, при данном уровне значимости гипотеза о том, что распределением исследуемого признака является распределение, Пуассона подтверждается.

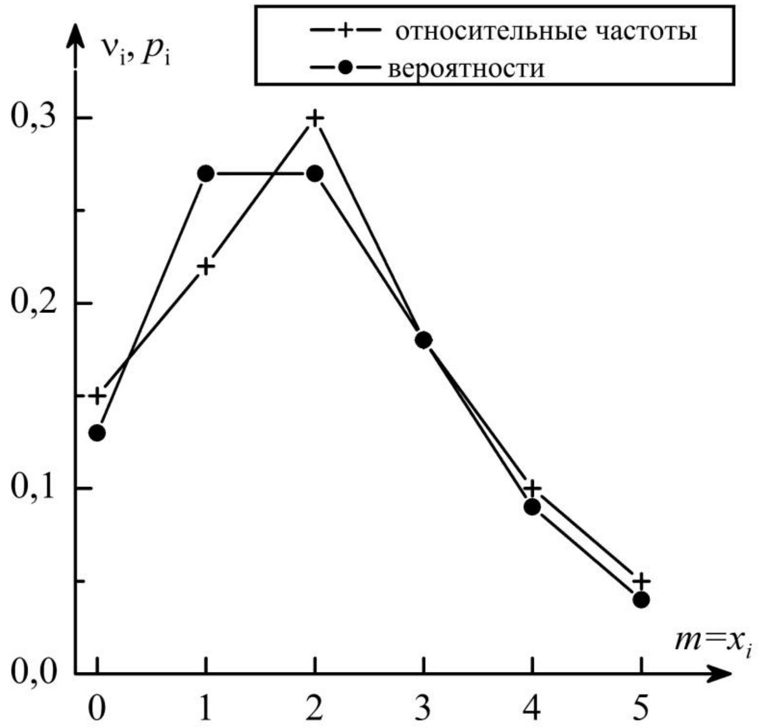


Рис.4.7. Полигоны эмпирического и теоретического распределений.

2. Проверить гипотезу при α=0,1 о том, что признак, представленный вариационным рядом из табл.4.8, распределен по нормальному закону.

Таблица 4.8

Доход x_i тыс.руб.	0-5	5-10	10-15	15-20	20-25	25-30	30-35
Число жителей m_i	31	96	158	240	270	157	48

Все данные, необходимые для вычислений представлены в таблице 3.8. Нам понадобится «представитель» каждого интервала $\overline{x}_i$, в качестве которого выбреем середину интервала. Используя данные таблицы, вычисляем выборочное среднее $\tilde{m}_x$ и выборочную дисперсию $\tilde{D}_x = \tilde{\sigma}_x^2$:

$$\tilde{m}_x = \frac{1}{1000}\sum_{i=1}^{7}\tilde{x}_i m_i\,,\ \tilde{D}_x = \frac{1}{1000}\sum_{i=1}^{7}\tilde{x}_i^2 m_i - \tilde{m}_x^2\,.$$

Так как объем выборки велик, n=1000, находим смещенную дисперсию, это несколько упрощает вычисления. Найденные выборочные значения используем как математическое ожидание и среднее квадратическое отклонение нормального распределения. В четвертой строке таблицы 4.9 находятся теоретические вероятности, которые вычисляем как

$$p_i = \Phi\left(\frac{b_i - \tilde{m}_x}{\tilde{\sigma}_x}\right) - \Phi\left(\frac{a_i - \tilde{m}_x}{\tilde{\sigma}_x}\right), \tag{4.21}$$

где a_i, b_i – левая и правая границы соответствующего интервала. В пятой строке таблицы записаны произведения np_i.

Таблица 4.9

Доход x_i тыс.руб.	0-5	5-10	10-15	15-20	20-25	25-30	30-35
$\bar{x}_i$	2,5	7,5	12,5	17,5	22,5	27,5	32,5
Число жителей m_i	31	96	158	240	270	157	48
p_i	0,022	0,081	0,185	0,267	0,241	0,137	0,049
np_i	22	81	185	267	241	137	49

Вычисленные выборочные параметры равны: $\tilde{m}_x \approx 18{,}9$, $\tilde{\sigma}_x \approx 7{,}19$.

Вычисленное по формуле (4.18б) значение критерия согласия равно $\chi^2 \approx 19{,}6$. Так как по наблюденным данным мы определяли *два* параметра распределения, число связей равно b=3, а число степеней свободы $f = 7 - 3 = 4$. По таблице находим критическое значение статистики, оно равно $\chi_c^2 = 7{,}78$. Так как $\chi^2 > \chi_c^2$ гипотеза о нормальном распределении признака отвергается. На рис. 4.8 показаны гистограмма статистического ряда 4.8 для относительных частот и плотность нормального распределения для вычисленных значений $\tilde{m}_x$, $\tilde{\sigma}_x$. Видим, что расхождение эмпирического и теоретического распределений действительно достаточно велико.

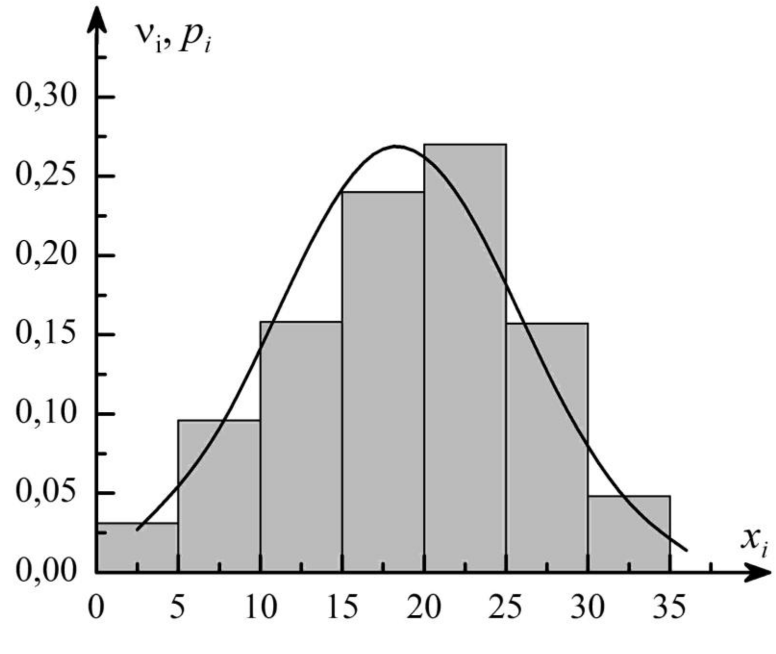


Рис.4.8. Гистограмма статистического ряда (4.8) и плотность нормального распределения (непрерывная кривая).

4.4. Корреляция и регрессия

Выборочный коэффициент корреляции. Пусть имеется выборка двух случайных величин. Обозначим эти величины как $x_i, i=1,2,...,n_1, y_j, j=1,2,...,n_2$. Для удобства запишем их в таблицу

Таблица 4.10

$x_i \backslash y_j$	y_1	y_2	...	y_l
x_1	m_{11}	m_{12}	...	m_{1l}
x_2	m_{21}	m_{22}	...	m_{2l}
…	....	...	...	...
x_k	m_{i1}	m_{i2}	...	m_{kl}

где m_{ij} – частоты, соответствующие парам значений (x_i, y_j).

Нам необходимо получить оценку коэффициента корреляции по выборочным значениям x_i, y_i, то есть выборочный коэффициент корреляции, который мы обозначим как $\tilde{r}_{xy}$. Обозначим сумму всех частот как *n*. В соответствии с общим принципом заменим в формулах (2.39б), (2.39в) математические ожидания средними арифметическими. Тогда в соответствии с (2.40) получим

$$\tilde{r}_{xy} = \frac{\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{l}(x_i - \tilde{m}_x)(y_j - \tilde{m}_y)m_{ij}}{\tilde{\sigma}_x\tilde{\sigma}_y} = \frac{\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{l}x_i y_j m_{ij} - \tilde{m}_x\tilde{m}_y}{\tilde{\sigma}_x\tilde{\sigma}_y}, \qquad (4.22)$$

где в знаменателе стоят выборочные среднеквадратичные отклонения (смещенные):

$$\tilde{\sigma}_x = \sqrt{\frac{1}{n}\sum_{i=1}^{k}x_i^2\sum_{j=1}^{l}m_{ij} - \tilde{m}_x^2}\ , \ \tilde{\sigma}_y = \sqrt{\frac{1}{n}\sum_{j=1}^{l}y_j^2\sum_{i=1}^{k}m_{ij} - \tilde{m}_y^2}\ .$$

Величину в числителе (4.22) следует рассматривать, конечно, как выборочный корреляционный момент

$$\tilde{K}_{xy} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{l}(x_i - \tilde{m}_x)(y_j - \tilde{m}_y)m_{ij} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{l}x_i y_j m_{ij} - \tilde{m}_x\tilde{m}_y\,. \qquad (4.23)$$

Метод наименьших квадратов.

Пусть имеются две выборки значений некоторых признаков *X, Y* генеральной совокупности. Для удобства запишем эти выборки в таблицу 4.11.

Таблица 4.11

x_i	x_1	x_2	x_3	...	x_n
y_i	y_1	y_2	y_3	...	y_n

Пусть у нас есть основания полагать, что значения x_i, y_i связаны некоторым неизвестным нам образом, т.е. y_i являются значениями некоторой неизвестной функции, вычисленной при значениях x_i ее аргумента. Процесс измерения всегда сопровождается сугубо случайными ошибками, поэтому точки (x_i, y_i) никогда не лягут на гладкую кривую. Задача заключается в том, чтобы подобрать гладкую кривую (функцию), которая наилучшим образом моделировала бы зависимость *y* от *x*. Фактически, мы предполагаем, что если бы не случайные (и неизбежные) ошибки измерения, эмпирические значения (x_i, y_i) легли бы на гладкую кривую, представляющую график некоторой функции $y = f(x)$. Выбор функции $f(x)$ не является задачей математической статистики, относительно этого выбора можно дать лишь общие рекомендации. Но любая функция содержит некоторое количество параметров. Например, линейная функции содержит два параметра *a, b*: $f(x) = a\,x + b$; функция, зависящая от *x* квадратичным образом, содержит в общем случае три параметра *a*, *b, c*: $f(x) = a\,x^2 + b\,x + c$; показательная функция может содержать два параметра: $y = ae^{bx}$ и т.д. Итак, будем полагать, что выбранная нами функция зависит от *m* параметров:

$$y = f(x; a_1, a_2, \ldots, a_m).$$

Задача математической статистики заключается в определении параметров $a_1, a_2, \ldots, a_m$, так чтобы теоретическая гладкая зависимость $f(x; a_1, a_2, \ldots, a_m)$ наилучшим образом описывала эмпирические точки (x_i, y_i). Слова «наилучшим образом» означают, что нам нужно так подобрать значения параметров $a_1, a_2, \ldots, a_m$, чтобы минимизировать сумму квадратов отклонений эмпирических значений y_i от значений соответствующей теоретической функции

$$S(a_1, a_2, \ldots, a_m) = \sum_{i=1}^{n} \left(y_i - f(x_i; a_1, a_2, \ldots, a_m)\right)^2 \to \min . \qquad (4.24)$$

В том случае, когда известны ошибки σ_i измерений величины *Y*, причем они различны, минимизируют выражение

$$S(a_1, a_2, \ldots, a_m) = \sum_{i=1}^{n} \frac{\left(y_i - f(x_i; a_1, a_2, \ldots, a_m)\right)^2}{\sigma_i^2} \to \min . \qquad (4.25)$$

Дело в том, что чем больше ошибка некоторого измерения, тем меньший вклад в сумму должно дать соответствующее слагаемое. Если все ошибки σ_i одинаковы, нет необходимости использовать (4.25) вместо (4.24).

Определение неизвестных параметров из минимизации выражения (4.24) (или (4.25)) составляет суть *метода наименьших квадратов*.

Для простоты рассмотрения мы полагаем далее, что все ошибки одинаковы.

Здесь мы подробно рассмотрим задачу аппроксимации эмпирических точек линейной функцией $f(x) = a\,x + b$:

$$S(a, b) = \sum_{i=1}^{n} \left(y_i - ax_i - b\right)^2 \to \min . \qquad (4.26)$$

Такая задача носит название линейной регрессии.

В точке минимума частные производные функции двух переменных $S(a,b)$ должны быть равны нулю

$$\begin{cases} \dfrac{\partial}{\partial a} S(a,b) = 0, \\ \dfrac{\partial}{\partial b} S(a,b) = 0. \end{cases} \qquad (4.27)$$

Распишем подробнее уравнения (4.27). Получим

$$\begin{cases} \sum\limits_{i=1}^{n}(y_i - ax_i - b)x_i = 0 \\ \sum\limits_{i=1}^{n}(y_i - ax_i - b) = 0 \end{cases}. \tag{4.28}$$

Или

$$\begin{cases} a\sum\limits_{i=1}^{n}x_i^2 + b\sum\limits_{i=1}^{n}x_i = \sum\limits_{i=1}^{n}x_i y_i \\ a\sum\limits_{i=1}^{n}x_i + bn = \sum\limits_{i=1}^{n}y_i \end{cases}. \tag{4.28a}$$

Таким образом, имеем систему двух линейных уравнений относительно двух неизвестных a, b. Решается система элементарно:

$$a = \frac{\frac{1}{n}\sum\limits_{i=1}^{n}x_i y_i - \frac{1}{n}\sum\limits_{i=1}^{n}x_i \frac{1}{n}\sum\limits_{i=1}^{n}y_i}{\frac{1}{n}\sum\limits_{i=1}^{n}x_i^2 - \left(\frac{1}{n}\sum\limits_{i=1}^{n}x_i\right)^2} = \frac{\frac{1}{n}\sum\limits_{i=1}^{n}x_i y_i - \tilde{m}_x\tilde{m}_y}{\tilde{\sigma}_x^2} = \frac{\tilde{\sigma}_y}{\tilde{\sigma}_x}\cdot\frac{\frac{1}{n}\sum\limits_{i=1}^{n}x_i y_i - \tilde{m}_x\tilde{m}_y}{\tilde{\sigma}_x\tilde{\sigma}_y}.$$

В последнем выражении стоит выборочный коэффициент корреляции для данных из таблицы 4.11:

$$\tilde{r}_{xy} = \frac{\frac{1}{n}\sum\limits_{i=1}^{n}x_i y_i - \tilde{m}_x\tilde{m}_y}{\tilde{\sigma}_x\tilde{\sigma}_y}.$$

Поэтому

$$a = \frac{\tilde{\sigma}_y}{\tilde{\sigma}_x}\tilde{r}_{xy}. \tag{4.29}$$

Для величины b несложно получить

$$b = \tilde{m}_y - \tilde{m}_x\frac{\tilde{\sigma}_y}{\tilde{\sigma}_x}\tilde{r}_{xy}. \tag{4.30}$$

Конечно, равенство нулю частных производных это необходимое, но не достаточное условие экстремума. Для проверки достаточного условия экстремума для функции двух переменных $S(a,b)$ следует вычислить величину

$$\Delta = S''_{aa}S''_{bb} - (S''_{ab})^2.$$

Если $\Delta > 0$ в предполагаемой точке экстремума, то экстремум действительно имеет место, причем, если $S''_{aa} > 0$, то данная точка является точкой минимума. Предлагаем проверить выполнение достаточного условия минимума самостоятельно.

Выпишем в окончательном виде уравнение линейной регрессии, т.е. уравнение прямой линии наилучшим образом описывающее эмпирические точки (x_i, y_i)

$$y = \frac{\tilde{\sigma}_y}{\tilde{\sigma}_x}\tilde{r}_{xy}\cdot x + \tilde{m}_y - \tilde{m}_x\frac{\tilde{\sigma}_y}{\tilde{\sigma}_x}\tilde{r}_{xy} \text{ или } y - \tilde{m}_y = \frac{\tilde{\sigma}_y}{\tilde{\sigma}_x}\tilde{r}_{xy}(x - \tilde{m}_x) \tag{4.31}$$

Можно получить и обратную зависимость, т.е. x от y. Здесь важно сказать, что в силу случайного характера связи величин X, Y, обратную зависимость ни в коем случае не следует определять, просто выражая x через y из (4.31). Необходимо в уравнении (4.31) поменять x на y, а y на x и учесть, что $\tilde{r}_{xy} = \tilde{r}_{yx}$:

$$x = \frac{\tilde{\sigma}_x}{\tilde{\sigma}_y} \tilde{r}_{xy} \cdot y + \tilde{m}_x - \tilde{m}_y \frac{\tilde{\sigma}_x}{\tilde{\sigma}_y} \tilde{r}_{xy} \text{ или } x - \tilde{m}_x = \frac{\tilde{\sigma}_x}{\tilde{\sigma}_y} \tilde{r}_{xy} (y - \tilde{m}_y). \qquad (4.31а)$$

Выборочный коэффициент корреляции величина случайная. Поэтому, если даже его абсолютная величина близка к единице, то это еще не говорит о сильной корреляционной зависимости, но может быть просто влиянием сугубо случайных факторов. Можно показать, что в предположении справедливости гипотезы H_0 *об отсутствии какой-либо корреляционной связи* величина

$$q = \frac{\tilde{r}_{xy} \sqrt{n-2}}{\sqrt{1 - \tilde{r}_{xy}^2}} \qquad (4.32)$$

подчиняется распределению Стьюдента с n–2 степенями свободы. Для достаточно больших выборок (n>50) распределение Стьюдента можно заменить нормальным распределением. Статистику (4.32) используют для проверки гипотезы о наличии сильной корреляционной связи. Именно, гипотеза H_0 отвергается если

$$|q| = \frac{|\tilde{r}_{xy}| \sqrt{n-2}}{\sqrt{1 - \tilde{r}_{xy}^2}} > q_c,$$

где q_c критическое значение, определяемое для заданного уровня значимости по таблице распределения Стьюдента или по таблице функции Лапласа.

Метод наименьших квадратов используют также для построения линий тренда. В статистике под трендом понимается закономерность в изменении какого-либо статистического признака во времени. В этом случае в роли величины x будет выступать неслучайная величина – время, но все уравнения останутся без изменений.

Примеры.

1. Имеются следующие данные об уровне механизации работ X (%) и производительности труда Y (изделий/сутки) для 10 однотипных предприятий (табл.4.14). Построить уравнения линейной регрессии $y(x)$, $x(y)$.

Таблица 4.12

x_i, %	32	36	41	47	54	55	56	60	67	69
y_i, 1/сутки	22	28	31	33	37	40	34	38	43	45

Вычисляем все необходимые величины:

$$\tilde{m}_x = \frac{1}{10} \sum_{i=1}^{10} x_i = 51{,}7, \quad \tilde{m}_y = \frac{1}{10} \sum_{i=1}^{10} y_i = 35{,}1,$$

$$\tilde{\sigma}_x^2 = \frac{1}{10} \sum_{i=1}^{10} x_i^2 - \tilde{m}_x^2 = 140{,}81, \quad \tilde{\sigma}_x \approx 11{,}87,$$

$$\tilde{\sigma}_y^2 = \frac{1}{10} \sum_{i=1}^{10} y_i^2 - \tilde{m}_y^2 = 44{,}09, \quad \tilde{\sigma}_y \approx 6{,}64,$$

$$\tilde{r}_{xy} = \frac{\frac{1}{10}\sum_{i=1}^{10} x_i y_i - \tilde{m}_x \tilde{m}_y}{\tilde{\sigma}_x \tilde{\sigma}_y} \approx 0{,}959\,.$$

Уравнения линейной регрессии: $y = 0{,}536x + 7{,}368$, $x = 1{,}713x - 8{,}429$.

Исходные данные и уравнение регрессии $y = a\,x + b$ показаны на рис. 4.9.

Коэффициент корреляции близок к единице, что может свидетельствовать о наличии сильной корреляционной связи между величинами *X*, *Y*.

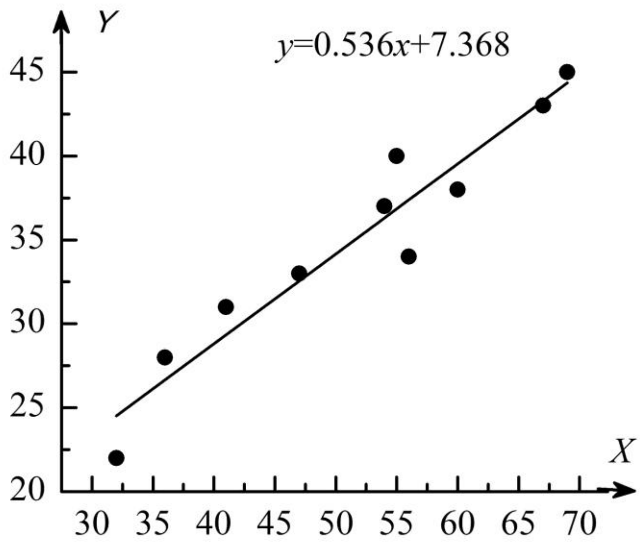


Рис.4.9. Исходные данные из табл. 4.12 (точки) и уравнение линейной регрессии (прямая линия).

Для уровня значимости α=0,05 проверим значимость выборочного коэффициента корреляции, используя статистику (4.32):

$$q = \frac{0{,}959\sqrt{10-2}}{\sqrt{1-0{,}959^2}} \approx 9{,}571\,.$$

По таблице распределения Стьюдента для числа степеней свободы 8 находим критическое значение статистики: q_c=2,306. Так как $q > q_c$, гипотеза H_0 об отсутствии корреляции отвергается, при данном уровне значимости вывод о наличии сильной корреляционной связи между величинами *X*, *Y* подтверждается.

2. В таблице 4.13 приведены данные: стаж работы рабочего на данном предприятии *X* (лет) и количество брака *Y* в процентах, допущенное рабочими за истекший период. Построить уравнения линейной регрессии $y(x)$, $x(y)$.

Таблица 4.13

Стаж X, лет	Середины интервалов	Допущенный брак Y, %						
		0-2	2-4	4-6	6-8	8-10		
	$\bar{y}_j$ \ $\bar{x}_i$	1	3	5	7	9	Сумма частот по строкам m_i	
0 – 5	2,5	1	3	10	7	2	23	5,52
5 – 10	7,5	2	7	8	2	1	20	4,30
10 – 15	12,5	7	8	6	1	0	22	3,09
15 – 20	17,5	12	4	2	0	0	18	1,89
20 – 25	22,5	12	4	0	1	0	17	1,82
	Сумма частот по столбцам, m_j	34	26	26	11	3		

Вычисляем все необходимые величины:

$$\tilde{m}_x = \frac{1}{100}\sum_{i=1}^{5}\bar{x}_i m_i = 11{,}8,\ \tilde{m}_y = \frac{1}{100}\sum_{j=1}^{5}\bar{y}_j m_j = 3{,}46,$$

$$\tilde{\sigma}_x^2 = \frac{1}{100}\sum_{i=1}^{5}\bar{x}_i^2 m_i - \tilde{m}_x^2 = 49{,}01,\ \tilde{\sigma}_x \approx 7{,}00,$$

$$\tilde{\sigma}_y^2 = \frac{1}{100}\sum_{j=1}^{5}\bar{y}_j^2 m_j - \tilde{m}_y^2 \approx 5{,}03,\ \tilde{\sigma}_y \approx 2{,}24,$$

$$\tilde{r}_{xy} = \frac{\frac{1}{100}\sum_{i=1}^{5}\sum_{j=1}^{5}\bar{x}_i\bar{y}_j m_{ij} - \tilde{m}_x\tilde{m}_y}{\tilde{\sigma}_x\tilde{\sigma}_y} \approx -0{,}623,$$

где m_{ij} – частоты, соответствующие парам значений $(\bar{x}_i, \bar{y}_j)$.

Из таблицы 3.14 видно, что общая тенденция такова, что с увеличением стажа работы процент брака уменьшается, соответственно и коэффициент корреляции получился отрицательным.

Теперь по формулам (4.31), (4.31а) получаем уравнения регрессии

$y = -0{,}200x + 5{,}81$, $x = -1{,}94y + 18{,}5$.

3. В таблице дана ожидаемая продолжительность жизни (мужчины) при рождении в России*

* Россия в цифрах. 2020: Крат.стат.сб./Росстат– М., 2020 – 550 с.

Таблица 4.14

Год	2000	2005	2010	2015	2016	2017	2018	2019
Год, t_i	0	5	10	15	16	17	18	19
Ожидаемая продолжит. жизни y_i, лет	59	58,9	63,1	65,9	66,5	67,5	67,8	68,2

Построить линейную функцию тренда.

При построении линии тренда удобно принять начальный (2000-й) год за нулевой. Во второй строке таблицы запишем количество лет, начиная с 2000-го года.

Вычисляем все необходимые величины:

$$\tilde{m}_t = \frac{1}{8}\sum_{i=1}^{8} t_i = 12{,}50,\ \tilde{m}_y = \frac{1}{8}\sum_{i=1}^{8} y_i = 64{,}61,$$

$$\tilde{\sigma}_t^2 = \frac{1}{8}\sum_{i=1}^{8} t_i^2 - \tilde{m}_t^2 = 41{,}25,\ \tilde{\sigma}_t \approx 6{,}42,$$

$$\tilde{\sigma}_y^2 = \frac{1}{8}\sum_{i=1}^{8} y_i^2 - \tilde{m}_y^2 \approx 12{,}88,\ \tilde{\sigma}_y \approx 3{,}59,$$

$$\tilde{r}_{ty} = \frac{\frac{1}{8}\sum_{i=1}^{8} t_i y_i - \tilde{m}_t \tilde{m}_y}{\tilde{\sigma}_t \tilde{\sigma}_y} \approx 0{,}979.$$

Уравнение линейной функции тренда: $y = 0{,}547t + 57{,}78$.

Корреляционное отношение. Как отмечалось выше, модуль коэффициента r_{xy} будет близок к единице, если величины X, Y не просто коррелированны, но функциональная связь между ними близка к линейной. Возможна ситуация, когда X, Y сильно коррелированы, но связь существенно нелинейная, в этом случае r_{xy} может быть мал по величине. Более универсальным показателем корреляции является корреляционное отношение ρ_{yx} или ρ_{xy} (в отличии от r_{xy} корреляционное отношение несимметрично по переменным x, y).

Сгруппируем выборочные значения признака X некоторым образом. Для каждой группы найдем среднее значение признака Y, которое обозначим как g_{iy}. Найдем величину

$$\delta_y^2 = \frac{1}{n}\sum_{i=1}^{k} (g_{iy} - \tilde{m}_y)^2 m_i, \qquad (4.33)$$

где k – число групп на которые мы разбили выборочные значения X, m_i – сумма частот по каждой группе. Величина δ_y^2 называется межгрупповой дисперсией. Корреляционным отношением Y по X называется

$$\rho_{yx} = \frac{\delta_y}{\tilde{\sigma}_y}. \qquad (4.34)$$

Аналогичным образом определяется корреляционное отношение ρ_{xy} (X по Y).

В том случае, когда корреляция между переменными отсутствует, внутригрупповое среднее g_{iy} будет близко к общему среднему $\tilde{m}_y$, величина δ_y^2 будет близка к нулю и, следовательно, будет близко к нулю и корреляционное отношение. С другой стороны, наибольшее значение межгрупповой дисперсии не может превышать общей выборочной дисперсии, поэтому корреляционное отношение не может превосходить единицы. Таким образом, корреляционное отношение – неотрицательная величина, не превосходящая единицы.

Найдем величину ρ_{yx} для данных таблицы 4.13. В крайнем правом столбце запишем средние значения g_{iy} переменной y в каждой строке, т.е. для каждой группы переменной x:

$$g_{iy} = \frac{1}{100}\sum_{j=1}^{5}\bar{y}_j m_{ij},\ i = 1,...,5.$$

Межгрупповая дисперсия

$$\delta_y^2 = \frac{1}{100}\sum_{i=1}^{5}(g_{iy} - \tilde{m}_y)^2 m_i \approx 2{,}05,\ \delta_y \approx 1{,}43$$

и

$$\rho_{yx} = \frac{1{,}43}{2{,}24} \approx 0{,}638.$$

Видим, что корреляционное отношение близко к найденному выше значению модуля коэффициента корреляции. Это значит, что зависимость величины Y от X действительно близка к линейной.

Задачи для самостоятельного решения.

1. В некотором регионе получены данные по количеству детей в семьях возраст родителей в которых не превышает 35 лет. Данные представлены вариационным рядом.

Таблица к задаче 1

Количество детей, x_i	0	1	2	3	4	5	6
Количество семей m_i	9	48	53	23	14	1	2

Для данного вариационного ряда построить полигон, определить среднее количество детей в расчете на одну семью.

2. В результате опроса 20 случайно выбранных студентов ВУЗа получены следующие баллы по дисциплине физическая культура: 2, 4, 3, 5, 2, 4, 4, 3, 4, 5, 3, 3, 3, 5, 4, 3, 2,5,4,4.

Построить дискретный вариационный ряд распределения студентов по баллам, изобразить его графически, найти средний балл.

3. Для контроля за точностью изготовления деталей, выполнены измерения 10 образцов и получены следующие значения диаметра X, мм: 15,2; 14,8; 15,0; 15,4; 16,0; 14,8; 15,1; 14,7; 15,6; 15,0.

Определить выборочную смещенную и несмещенную дисперсии для диаметра.

4. Для вариационного ряда заданного таблицей 4.2 найти доверительный интервал, в который математическое ожидания уровня дохода попадает с вероятностью 0,95. Определить, каким должен быть объем выборки, чтобы те же границы обеспечивались с вероятностью 0,9973 (дисперсию признака считать неизменной)

5. Для ста предприятий некоторой отрасли собраны данные по годовому объему продукции по отношению к предыдущему году в %. Данные представлены в таблице.

Для данного вариационного ряда построить гистограмму и кумуляту, определить интервал, в который объем продукции попадает с вероятностью 0.99.

Таблица к задаче 5

Годовой объем, x%	86-90	90-94	94-98	98-102	102-106	106-110	110-114
Количество предприятий, m	6	10	8	24	34	14	4

6. В результате выборочного обследования малых торговых предприятий региона, получены следующие данные:

Таблица к задаче 6

Номер предприятия	Товарооборот, млн.руб./мес.	Номер предприятия	Товарооборот, млн.руб./мес.	Номер предприятия	Товарооборот, млн.руб./мес.
1	4,3	13	4,5	25	5,1
2	6,9	14	5,9	26	6,1
3	6,0	15	5,9	27	5,3
4	5,0	16	6,8	28	7,5
5	5,2	17	5,8	29	4,7
6	5,5	18	4,0	30	4,7
7	2,8	19	6,6	31	5,4
8	6,1	20	5,8	32	4,2
9	4,9	21	6,6	33	7,0
10	5,3	22	4,5	34	5,9
11	4,4	23	5,3	35	3,5
12	5,5	24	4,6	36	4,4

а) Построить интервальный вариационный ряд, число интервалов определить по правилу Стерджеса;

б) построить гистограмму и кумулятивную кривую;

в) определить интервал, в который средний товарооборот попадает с вероятностью 0.99.

7. В результате выборочного обследования получены следующие данные для 60 предприятий области

Таблица к задаче 7

№ предп.	Годовая прибыль, млн.руб.	№ предп.	Годовая прибыль, млн.руб.	№ предп.	Годовая прибыль, млн.руб.
1	20,7	21	39,0	41	24,9
2	15,1	22	22,1	42	20,1
3	15,9	23	23,0	43	13,1
4	25,5	24	28,9	44	17,4
5	19,0	25	23,5	45	29,3
6	27,1	26	35,5	46	36,0
7	31,1	27	28,6	47	22,0
8	16,1	28	28,8	48	28,9
9	19,8	29	34,9	49	34,3
10	34,3	30	21,0	50	23,6
11	33,2	31	24,1	51	37,1
12	16,4	32	29,0	52	24,0
13	10,2	33	32,5	53	33,0
14	18,0	34	36,8	54	24,1
15	32,3	35	28,8	55	20,2
16	14,0	36	39,5	56	14,5
17	24,5	37	35,4	56	26,4
18	30,7	38	25,3	58	23,0
19	41,0	39	45,0	59	18,0
20	27,8	40	28,3	60	17,2

а) Построить интервальный вариационный ряд , число интервалов взять равным 7;
б) построить гистограмму и кумулятивную кривую;
в) определить интервал, в который значение вариационного признака попадает с вероятностью 0,99.

8. Найти доверительный интервал для оценки с вероятностью $\alpha = 0{,}95$ неизвестного математического ожидания m_x нормально распределенного признака X генеральной совокупности, если выборочное среднее равно 14, несмещенная оценка среднего квадратического отклонения равна 6, объем выборки равен 25.

а) Выполнить расчет, используя нормальное распределение для выборочного среднего;

б) использовать распределение Стьюдента.

9. Десять независимых измерений признака X генеральной совокупности дали следующие значения: 1,61; 1,65; 1,53; 1,7; 1,9; 1,9; 1,75; 1,87; 1,39; 1,7. Найти интервал, в который математическое ожидание признака попадает с вероятностью 0,99.

10. Опрос 60 жителей города показал, что 36 человек будут голосовать за действующего мэра на предстоящих выборах. Найти границы, в которых с вероятностью 0,9 заключена доля граждан в %, которые будут поддерживать на предстоящих выборах действующего мэра.

11. Решить ту же задачу, но а) теперь опросили 300 человек, 180 будут поддерживать мэра. б) Опросили 3000 человек, 1800 будут поддерживать мэра.

12. На станке в автоматическом режиме изготавливаются металлические втулки. Проектная длина втулки должна составлять 52 мм. Однако для выборки из 150 втулок длина

составил 52,2 мм. При среднем квадратическом отклонении 1,2 мм и при уровне значимости $\alpha=0{,}01$ проверить гипотезу H_0 том, что станок не требует подналадки.

13. Для десяти предприятий отрасли получен следующий процент брака готовой продукции: 1,6; 1,4; 1,0; 1,8; 0,9; 2,5; 3,4; 0,8; 1,8; 2,8. Для уровня значимости 10% проверить гипотезу H_0 о том, что средний уровень брака не превышает одного процента.

14. Игральный кубик подброшен 120 раз. Шестерка выпала 15 раз. Можно ли на уровне значимости $\alpha=0{,}1$ принять гипотезу H_0: вероятность p появления шестерки равна 1/6 (конкурирующая гипотеза H_1: $p \neq 1/6$)? *Указание:* для дисперсии частоты использовать выражение для дисперсии распределения Бернулли при $p=1/6$.

15. Предыдущая задача, но игральный кубик подброшен 480 раз, шестерка выпала 60 раз.

16. На двух станках в автоматическом режиме изготавливаются металлические втулки. По выборке объема $n=120$ для первого станка средний диаметр втулки составил 25,1 мм. По выборке $n=100$ для второго станка средний диаметр составил 19,8 мм. Выборочные дисперсии равны соответственно 1,44 мм^2 и 1,40 мм^2. При уровне значимости $\alpha=0{,}05$ проверить гипотезу H_0 о равенстве генеральных средних для станков при конкурирующей гипотезе H_1: $m_x \neq m_y$

17. Процент брака для продукции первого цеха, найденный по выборке $n=320$ составил 2,5%, для второго цеха он составил 1,6% при $n=250$. При уровне значимости $\alpha=0{,}05$ проверить гипотезу H_0 о равенстве процентов брака для первого и второго цеха, при конкурирующей гипотезе, что проценты брака различны.

18. Для первой бригады при выборке объема $n=700$ количество бракованных изделий оказалось равным 70, для второй бригады при выборке $n=800$ количество бракованных изделий равно 60. Можно ли при уровне значимости $\alpha=0{,}1$ утверждать, что процент брака одинаков для обеих бригад при конкурирующем утверждении, что он выше для первой бригады?

19. На двух станках изготовляются втулки. Для первого станка по выборке объема $n=10$ выборочная дисперсия отклонения длины втулки от проектной равна 0,9 мм^2, для второго станка по выборке $n=12$ выборочная дисперсия отклонения равной 1,5 мм^2. Для уровня значимости $\alpha=0{,}01$ проверить гипотезу H_0 о равенстве генеральных дисперсий при конкурирующей гипотезе H_1: дисперсия выше для второго станка.

20. Для пятидесяти станций скорой помощи получена следующая статистика числа X ложных вызовов в течении суток

Таблица к задаче 20

Количество ложных вызовов x_i	0	1	2	3	4	5	6	7
Количество станций m_i	4	7	16	13	6	2	1	1

С уровнем значимости $\alpha=0{,}1$ проверить гипотезу H_0, что для признака X имеет место распределение Пуассона. Построить полигон эмпирического и теоретического распределений. *Указание*: в качестве параметра a взять среднее арифметическое выборочного среднего и дисперсии, при вычислении χ^2 три последних разряда объединить в один.

21. За истекший месяц для сорока магазинов получены следующие данные по количеству X выявленных нарушений правил торговли

Таблица к задаче 21

Количество нарушений x_i	0	1	2	3	4
Количество магазинов m_i	14	12	5	5	4

С уровнем значимости α=0,05 проверить гипотезу H_0, что для признака X имеет место распределение Пуассона. Построить полигон эмпирического и теоретического распределений. *Указание*: в качестве параметра a взять среднее арифметическое выборочного среднего и дисперсии.

22. В таблице для выборки из сорока избирательных участков приведены значения признака X – разницы количества избирателей, проголосовавших за кандидата (в процентном выражении), и рейтинга, полученного по опросу избирателей до голосования.

а) Построить интервальный статистический ряд с числом интервалов k=6 (k выбрано в соответствии с правилом Стерджеса);

б) для интервального ряда построить гистограмму;

в) с уровнем значимости α=0,1 проверить гипотезу H_0: генеральное среднее признака X равно нулю;

г) с уровнем значимости α=0,05 проверить гипотезу H_0 о нормальном распределении признака X.

Таблица к задаче 22

Номер участка	Значение признака X,%	Номер участка	Значение признака X,%	Номер участка	Значение признака X,%
1	0,8	15	9,8	29	–5,5
2	10,0	16	12,0	30	14,7
3	–4,2	17	–2,2	31	12,4
4	0,2	18	–0,8	32	2,8
5	0,5	19	3,5	33	–3,5
6	–3,6	20	–3,6	34	13,5
7	10,2	21	0,6	35	2,0
8	1,7	22	2,7	36	0,3
9	1,8	23	–2,2	37	–2,0
10	–4,0	24	–4,5	38	0,0
11	0,2	25	4,0	39	9,5
12	5,0	26	5,0	40	10,0
13	–6,2	27	–6,2		
14	6,7	28	–9,3		

23. Для данных задачи 7 проверить гипотезу H_0 о том, что рассматриваемый признак X (годовая прибыль предприятия, млн.руб) распределен по нормальному закону. Уровень значимости α принять равным 0,1.

24. Имеются следующие данные об энерговооруженности труда X (кВт) в расчете на одного работающего и производительности труда Y (тыс. руб/сутки) для 10 однотипных предприятий. Найти коэффициент корреляции, построить уравнения линейной регрессии $y(x)$, $x(y)$, проверить гипотезу о сильной корреляционной связи между величинами X, Y при уровне значимости α=0,05.

Таблица к задаче 24.

x_i	3,2	3,7	4,0	4,8	6,0	5,4	5,2	5,4	6,0	9,0
y_i	8,4	8,8	9,1	9,8	10,6	10,7	11,1	11,8	12,1	12,4

25. В таблице приведены коэффициенты рождаемости в Российской Федерации[*].

Таблица к задаче 25

Год	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
Число родившихся на 1000 чел.	9,0	9,7	10,2	10,4	10,2	10,4	11,3	11,4	11,8	12,0	12,0	12,8

а) Найти линейную функцию тренда (2000-й год примите за нулевой);

б) На основе тренда дать прогноз числа родившихся (промилле) в 2019 г. и сравнить его с имеющимися статистическими данными.

26. В таблице для выборки из ста студентов ВУЗа представлены данные по посещаемости X (количество пропущенных занятий, %) и успеваемости Y (количество баллов за семестр по 100-бальной системе). Найти коэффициент корреляции r_{xy}. С уровнем значимости 0,05 проверить гипотезу о сильной корреляционной связи величин Y, X.

Таблица к задаче 26.

Пропуски X, %	Середины интервалов	Количество баллов Y				
		0–20	20–40	40–60	60–80	80–100
	$\bar{y}_j$ / $\bar{x}_i$	10	30	50	70	90

[*] Статистика. Под. ред. И.И.Елисеевой. М. Юрайт.2015.– 446 стр.

0 - 15	7,5	0	0	1	11	15
15 - 30	22,5	0	1	2	12	8
30 - 45	37,5	5	4	4	1	0
45 - 60	52,5	8	3	1	2	0
60 - 75	67,5	8	2	1	0	0
75-100	87,5	10	1	0	0	0

27. В таблице приведены данные: суточная выработка продукции Y (тонны) и величина основных производственных фондов X (млн. руб.) для совокупности 50 однотипных предприятий.

Таблица к задаче 27.

Величина ОФП, млн. руб.	Середины интервалов	Суточная выработка продукции, тонны				
		7-11	11-15	15-19	19-23	23-27
	y_i / x_i	9	13	17	21	25
20 – 24	22	2	1	0	0	0
24 – 28	26	3	6	4	0	0
28 – 32	30	0	3	11	7	0
32 – 36	34	0	1	2	6	2
36 – 40	38	0	0	0	1	1

а) Найти выборочный коэффициент корреляции;

б) построить уравнения линейной регрессии $y(x)$, $x(y)$;

в) найти корреляционное отношение ρ_{yx}.

28. По выборке объема n=18 получены уравнения регрессии $y = -1{,}31x + 28{,}1$, $x = -0{,}580y + 18{,}7$. Определить коэффициент корреляции и проверить его значимость при уровне значимости α=0,01.

Ответы: 1) 1,97. 2) 3,6 3) Смещенная оценка дисперсии $\approx$0,148 мм^2, несмещенная$\approx$ 0,165 мм^2. 4) (18,03; 18,96); $n\approx$2343. 5) (99,6; 102,7). 6) в) (4,96; 5,84). 7) в) (23,4; 28,6) 8) а) (11,65; 16,35) б) (11,52; 16,48). 9) (1,51; 1,89) 10) (49,6%; 70,4%). 11) а) (55,3%; 64,7%) б) (58,5%; 61,5%). 12) Область принятия гипотезы H_0 (доверительный интервал) (51,75; 52,25), гипотеза H_0 принимается. 13) Гипотеза H_0 отвергается (критическое значение статистики 0,7, найденное значение 0,8). 14) при данном уровне значимости гипотеза H_0 принимается. 15) при данном уровне значимости гипотеза H_0 должна быть отвергнута. 16) Гипотеза принимается (критическое значение статистики 1,96, найденное значение 1,86). 17) Гипотеза H_0 принимается (найденное значение критерия 0,74, критическое значение статистики 1,96). 18) при данном уровне значимости принимается конкурирующая гипотеза (найденное значение критерия 1,72, критическое значение 1,28) 19) при данном уровне значимости гипотеза H_0 принимается (найденное значение статистики 1,67, критическое значение 5,78). 20) при данном уровне значимости гипотеза H_0 принимается (найденное значение χ^2=3,57, критическое значение 7,78). 21) при данном уровне значимости гипотеза H_0 должна быть отвергнута (найденное значение χ^2=8,38, критическое значение 7,82). 22) в) Область принятия гипотезы H_0 (доверительный интервал) (–1,55; 1,55), гипотеза H_0 отвергается; г) гипотеза H_0 принимается (найденное значение χ^2=3,40, критическое значение статистики 7,82). 23) гипотеза H_0 принимается (найденное значение χ^2=1,66, критическое значение статистики 9,49). 24) $r_{xy} = 0{,}854$, $y = 0{,}746x + 6{,}549$, $x = 0{,}977y - 4{,}970$, гипотеза о сильной корреляции подтверждается (найденное значение критерия 4,64 при критическом значении 2,31). 25) $y = 0{,}303t + 8{,}964$, прогнозируемое количество родившихся в 2019 г. 14,7‰. 26) $r_{xy} = -0{,}827$. Отвергается гипотеза H_0 об отсутствии корреляционной связи, тем самым подтверждается конкурирующая гипотеза: наличие сильной корреляционной связи (значение критерия 14,8 при критическом значении статистики 1,96). 27) а) r_{xy}=0,740 б) $y = 0{,}845x - 8{,}17$, $x = 0{,}648y + 18{,}7$ в) ρ_{yx}=0,750. 28) $r_{xy} \approx -0{,}872$, значимость подтверждается (найденное значение критерия 7,11 при критическом значении 2,92).

Библиографический список

1. Вентцель Е.С. Теория вероятностей: учебник для вузов /Е.С.Вентцель. – 10-е изд., стер.–М.: Высш.шк.,2006.-575 с.
2. Вентцель Е.С., Овчаров Л.А. Теория вероятностей и ее инженерные приложения. Учеб. пособие для втузов/ 2-е изд., стер.– М.: Высш. шк., 2000.– 480 с.
3. Гнеденко Б.В. Курс теории вероятностей : Учебник /Изд 10-е, доп. – М.: Книжный дом "ЛИБРОКОМ", 2011.–488 с.
4. Гмурман В.Е. Теория вероятностей и математическая статистика: Учеб. пособие для вузов/В.Е. Гмурман. - 9-е изд., стер.– М.: Высш.шк., 2003.–479 с.
5. Кремер Н. Ш. Теория вероятностей и математическая статистика : учебник и практикум для вузов / Н. Ш. Кремер. – 5-е изд., перераб. и доп. – Москва : Издательство Юрайт, 2019. – 538 с.

Приложения

Приложение 1

Таблица функции Лапласа $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-t^2/2} dt$. Крайний левый столбец значения аргумента x с точностью до десятых, верхняя строка – сотые доли аргумента.

x	0	1	2	3	4	5	6	7	8	9
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	0,41308	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	0,49996									
4,5	0,49997									
5,0	0,49999									

Приложение 2

Распределение Стьюдента: $2\int_0^{x_\beta} S_{n-1}(t)dt = \beta,\ f = n-1$.

f \ β	0.80	0.90	0.95	0.98	0.99	0.995	0.998	0.999
1	3.0770	6.3130	12.7060	31.820	63.656	127.656	318.306	636.619
2	1.8850	2.9200	4.3020	6.964	9.924	14.089	22.327	31.599
3	1.6377	2.35340	3.182	4.540	5.840	7.458	10.214	12.924
4	1.5332	2.13180	2.776	3.746	4.604	5.597	7.173	8.610
5	1.4759	2.01500	2.570	3.649	4.0321	4.773	5.893	6.863
6	1.4390	1.943	2.4460	3.1420	3.7070	4.316	5.2070	5.958
7	1.4149	1.8946	2.3646	2.998	3.4995	4.2293	4.785	5.4079
8	1.3968	1.8596	2.3060	2.8965	3.3554	3.832	4.5008	5.0413
9	1.3830	1.8331	2.2622	2.8214	3.2498	3.6897	4.2968	4.780
10	1.3720	1.8125	2.2281	2.7638	3.1693	3.5814	4.1437	4.5869
11	1.363	1.795	2.201	2.718	3.105	3.496	4.024	4.437
12	1.3562	1.7823	2.1788	2.6810	3.0845	3.4284	3.929	4.178
13	1.3502	1.7709	2.1604	2.6503	3.1123	3.3725	3.852	4.220
14	1.3450	1.7613	2.1448	2.6245	2.976	3.3257	3.787	4.140
15	1.3406	1.7530	2.1314	2.6025	2.9467	3.2860	3.732	4.072
16	1.3360	1.7450	2.1190	2.5830	2.9200	3.2520	3.6860	4.0150
17	1.3334	1.7396	2.1098	2.5668	2.8982	3.2224	3.6458	3.965
18	1.3304	1.7341	2.1009	2.5514	2.8784	3.1966	3.6105	3.9216
19	1.3277	1.7291	2.0930	2.5395	2.8609	3.1737	3.5794	3.8834
20	1.3253	1.7247	2.08600	2.5280	2.8453	3.1534	3.5518	3.8495
21	1.3230	1.7200	2. 0790	2.5170	2.8310	3.1350	3.5270	3.8190
22	1.3212	1.7117	2.0739	2.5083	2.8188	3.1188	3.5050	3.7921
23	1.3195	1.7139	2.0687	2.4999	2.8073	3.1040	3.4850	3.7676
24	1.3178	1.7109	2.0639	2.4922	2.7969	3.0905	3.4668	3.7454
25	1.3163	1.7081	2.0595	2.4851	2.7874	3.0782	3.4502	3.7251
26	1.315	1.705	2.059	2.478	2.778	3.0660	3.4360	3.7060
27	1.3137	1.7033	2.0518	2.4727	2.7707	3.0565	3.4210	3.6896
28	1.3125	1.7011	2.0484	2.4671	2.7633	3.0469	3.4082	3.6739
29	1.3114	1.6991	2.0452	2.4620	2.7564	3.0360	3.3962	3.8494
30	1.3104	1.6973	2.0423	2.4573	2.7500	3.0298	3.3852	3.6460
32	1.3080	1.6930	2.0360	2.4480	2.7380	3.0140	3.3650	3.6210
34	1.3070	1.6909	2.0322	2.4411	2.7284	3.9520	3.3479	3.6007
36	1.3050	1.6883	2.0281	2.4345	2.7195	9.490	3.3326	3.5821
38	1.3042	1.6860	2.0244	2.4286	2.7116	3.9808	3.3190	3.5657
40	1.303	1.6839	2.0211	2.4233	2.7045	3.9712	3.3069	3.5510
42	1.320	1.682	2.018	2.418	2.6980	2.6930	3.2960	3.5370
44	1.301	1.6802	2.0154	2.4141	2.6923	3.9555	3.2861	3.5258
46	1.300	1.6767	2.0129	2.4102	2.6870	3.9488	3.2771	3.5150
48	1.299	1.6772	2.0106	2.4056	2.6822	3.9426	3.2689	3.5051
50	1.298	1.6759	2.0086	2.4033	2.6778	3.9370	3.2614	3.4060

Приложение 3

Значение величины χ^2 в зависимости от α и f

α / f	0,99	0,98	0,95	0,90	0,80	0,70	0,50	0,30	0,20	0,10	0,05	0,02	0,01	0,001
1	0,000	0,001	0,004	0,016	0,064	0,148	0,455	1,074	1,642	2,71	3,84	5,41	6,64	10,83
2	0,020	0,040	0,103	0,211	0,446	0,713	1,386	2,41	3,22	4,60	5,99	7,82	9,21	13,82
3	0,115	0,185	0,352	0,584	1,005	1,424	2,37	3,66	4,64	6,25	7,82	9,84	11,34	16,27
4	0,297	0,429	0,711	1,064	1,649	2,20	3,36	4,88	5,99	7,78	9,49	11,67	13,28	18,46
5	0,554	0,752	1,145	1,610	2,34	3,00	4,35	6,06	7,29	9,24	11,07	13,39	15,09	20,5
6	0,872	1,134	1,635	2,20	3,07	3,83	5,35	7,23	8,56	10,64	12,59	15,03	16,81	22,5
7	1,239	1,564	2,17	2,83	3,82	4,67	6,35	838	9,80	12,02	14,07	16,62	18,48	24,3
8	1,646	2,03	2,73	3,49	4,59	5,53	7,34	9,52	11,03	13,36	15,51	18,17	20,1	26,1
9	2,09	2,53	3,32	4,17	5,38	6,39	8,34	10,66	12,24	14,68	16,92	19,68	21,7	27,9
10	2,56	3,06	3,94	4,86	6,18	7,27	9,34	11,78	13,44	15,99	18,31	21,2	23,2	29,6
11	3,05	3,61	4,58	5,58	6,99	8,15	10,34	12,90	14,63	17,28	19,68	22,6	24,7	31,3
12	3,57	4,18	5,23	6,30	7,81	9,03	11,34	14,01	15,81	18,55	21,0	24,1	26,2	32,9
13	4,11	4,76	5,89	7,04	8,63	0,93	12,34	15,12	16,98	19,81	22,4	25,5	27,7	34,6
14	4,66	5,37	6,57	7,79	9,47	10,82	13,34	16,22	18,15	21,1	23,7	26,9	29,1	36,1
15	5,23	5,98	7,26	8,55	10,31	11,72	14,34	17,32	19,31	22,3	25,0	28,3	30,6	37,7
16	5,81	6,61	7,96	9,31	11,15	12,62	15,34	18,42	20,5	23,5	26,3	29,6	32,0	39,3
17	6,41	7,26	8,67	10,08	12,00	13,53	16 34	19,51	21,6	24,8	27,6	31,0	33,4	40,8
18	7,02	7,91	9,39	10,86	12,86	14,44	17,34	20,6	22,8	26,0	28,9	32,3	34,8	42,3
19	7,63	8,57	10,11	11,65	13,72	15,35	18,34	21,7	23,9	27,2	30,1	33,7	36,2	43,8
20	8,26	9,24	10,85	12,44	14,58	16,27	19,34	22,8	25,0	28,4	31,4	35,0	37,6	45,3
21	8,90	9,92	11,59	13,24	15,44	17,18	20,3	23,9	26,2	29,6	32,7	36,3	38,9	46,8
22	9,54	10,60	12,34	14,04	16,31	18,10	21,3	24,0	27,3	30,8	33,9	37,7	40,3	48,3
23	10,20	11,29	13,09	14,85	17,19	19,02	22,3	26,0	28,4	32,0	35,2	39,0	41,6	49,7
24	10,86	11,99	13,85	15,66	18,06	19,94	23,3	27,1	29,6	33,2	36,4	40,3	43,0	51,2
25	11,52	12,70	14,61	16,47	18,94	20,9	24,3	23,2	30,7	34,4	37,7	41,7	44,3	52,6
26	2,20	13,41	15,38	17,29	19,82	21,8	25,3	29,2	31,8	35,6	38,9	42,9	45,6	54,1
27	12,88	14,12	16,15	18,11	20,7	22,7	26,3	30,3	32,9	36,7	40,1	44,1	47,0	55,5
28	13,56	14,85	16,93	18,94	21,6	23,6	27,3	31,4	34,0	37,9	41,3	45,4	48,3	56,9
29	14,26	15,57	17,71	19,77	22,5	24,6	28,3	32,5	35,1	39,1	42,6	46,7	49,6	58,3
30	14,95	16,31	18,19	20,6	23,4	25,5	29,3	33,5	36,2	40,3	43,8	48,0	50,9	59,7

Учебное издание

С.Н. Фадеев, И.В. Зайцева

ТЕОРИЯ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Печатается в авторской редакции.

Подписано в печать 30.12.2021. Формат 60×90 1/8. Гарнитура Times New Roman.

Печать цифровая. Усл. печ. л. 11,75. Тираж 25 экз. Заказ № 1176.

РГГМУ, 192007, Санкт-Петербург, Воронежская ул., д. 79.