Министерство науки и высшего образования Российской Федерации

Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Сибирский государственный университет геосистем и технологий»
(СГУГиТ)

В. Л. Неклюдова, В. П. Вербная

ТЕОРИЯ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Утверждено редакционно-издательским советом университета
в качестве учебного пособия для обучающихся
по направлению подготовки 10.03.01 Информационная безопасность
(уровень бакалавриата)

Новосибирск
СГУГиТ
2024

УДК 519.2
Н476

Рецензенты: кандидат физико-математических наук, старший научный сотрудник института математики им С. Л. Соболева СО РАН *С. Е. Хрущев*

кандидат физико-математических наук, доцент, СГУГиТ *А. В. Логачев*

Неклюдова, В. Л.

Н476 Теория вероятностей и математическая статистика : учебное пособие / В. Л. Неклюдова, В. П. Вербная. – Новосибирск : СГУГиТ, 2024. – 92 с. – Текст : непосредственный.

ISBN 978-5-907711-62-4

Учебное пособие подготовлено кандидатом физико-математических наук, доцентом В. Л. Неклюдовой, старшим преподавателем В. П. Вербной на кафедре высшей математики СГУГиТ.

Пособие содержит теоретический материал по дисциплине «Теория вероятностей и математическая статистика» и включает такие темы, как случайные события, случайные величины, основные законы распределения, точечное и интервальное оценивание параметров распределения, проверка статистических гипотез.

Учебное пособие предназначено для обучающихся по направлению подготовки 10.03.01 Информационная безопасность (уровень бакалавриата), а также может быть использовано в организации учебного процесса по данной дисциплине по направлениям подготовки 05.03.03 Картография и геоинформатика, 12.03.01 Приборостроение, 12.03.02 Оптотехника, 12.03.03 Фотоника и оптоинформатика, 21.03.02 Землеустройство и кадастры, 27.03.01 Стандартизация и метрология, 27.03.05 Инноватика.

Рекомендовано к изданию кафедрой высшей математики, Ученым советом Института геодезии и менеджмента СГУГиТ.

Печатается по решению редакционно-издательского совета СГУГиТ

УДК 519.2

ISBN 978-5-907711-62-4

© СГУГиТ, 2024

ОГЛАВЛЕНИЕ

Введение 5

1. Случайные события 6
   1.1. Понятие события 6
   1.2. Вероятность 7
   1.3. Элементы комбинаторики 9
   1.4. Операции над событиями 12
   1.5. Теоремы сложения вероятностей 13
   1.6. Теоремы умножения вероятностей 13
   1.7. Формула полной вероятности. Формула Байеса 15

2. Случайные величины 18
   2.1. Понятие случайной величины. Дискретные случайные величины 18
   2.2. Числовые характеристики дискретной случайной величины 22
   2.3. Функция распределения 25
   2.4. Непрерывные случайные величины 27
   2.5. Числовые характеристики непрерывной случайной величины 31
   2.6. Основные законы распределения 32

3. Повторные независимые испытания 44
   3.1. Схема Бернулли 44
   3.2. Формула Пуассона 44
   3.3. Формулы Муавра – Лапласа 45

4. Системы двух случайных величин 48
   4.1. Дискретные двумерные случайные величины 48
   4.2. Условные законы распределения 50
   4.3. Ковариация и коэффициент корреляции 50

5. Основные понятия математической статистики 56
   5.1. Выборка и вариационный ряд 56
   5.2. Теоретическая и эмпирическая функция распределения 58
   5.3. Числовые характеристики выборки 60

6. Точечное оценивание параметров распределения 63
   6.1. Статистические оценки и их свойства 63
   6.2. Оценки математического ожидания и дисперсии 64
   6.3. Оценка параметров распределения методом моментов 65

7. Основные распределения математической статистики 68
8. Интервальное оценивание параметров распределения..... 71
8.1. Понятие доверительного интервала 71
8.2. Свойства выборок из нормального распределения..... 72
8.3. Доверительные интервалы для параметров нормального распределения 72
9. Корреляционный анализ..... 77
9.1. Оценка коэффициента корреляции 77
9.2. Уравнение линейной регрессии..... 78
10. Проверка статистических гипотез 81
10.1. Статистические гипотезы и их проверка 81
10.2. Проверка гипотезы о равенстве дисперсии нормально распределенной генеральной совокупности некоторому значению..... 82
10.3. Проверка гипотезы о равенстве математического ожидания нормально распределенной генеральной совокупности некоторому значению 84
10.4. Проверка гипотезы о доле признака..... 86
10.5. Проверка гипотезы о равенстве средних двух нормально распределенных генеральных совокупностей 87
Заключение 90
Библиографический список..... 91

ВВЕДЕНИЕ

Теория вероятностей изучает случайные явления и их закономерности и выступает теоретической базой для математической статистики. Математическая статистика, в свою очередь, позволяет выявлять и исследовать взаимосвязи между явлениями и процессами в самых разных областях науки и практики [1–3, 5].

Теория вероятностей и математическая статистика являются важнейшей частью математического аппарата современной науки и широко используются при моделировании физических, биологических, социальных процессов. В области компьютерных наук методы теории вероятностей и математической статистики применяются при исследовании алгоритмов шифрования и дешифровки, глубинном анализе данных, обучении нейросетей, организации систем распределенных вычислений.

Настоящее издание предназначено для обучающихся по направлению подготовки 10.03.01 Информационная безопасность (уровень бакалавриата) в качестве учебного пособия по дисциплине «Теория вероятностей и математическая статистика» и содержит теоретические сведения по всем разделам курса дисциплины, сопровождаемые многочисленными примерами.

Учебное пособие является частью учебно-методического комплекса дисциплины, включающего в себя также сборник задач, разработанный теми же авторами [4], содержащий материалы для практический занятий и самостоятельной работы обучающихся, а также таблицы распределений.

Данное учебное пособие в комплексе со сборником задач могут использоваться обучающимися по другим направлениям подготовки.

1. СЛУЧАЙНЫЕ СОБЫТИЯ

1.1. Понятие события

Теория вероятностей изучает явления, которые нельзя полностью предсказать, – ***случайные явления***, а также закономерности в такого рода явлениях и процессах. Базовыми понятиями в теории вероятностей являются событие и эксперимент.

Под ***экспериментом*** (***испытанием***) будем понимать выполнение комплекса условий и действий, при котором имеет место некоторый результат, который невозможно точно предугадать. Этот результат называется ***событием*** (или ***случайным событием***).

Пример 1. Примером эксперимента может выступать бросание кубика (игральной кости). Результат эксперимента определяется числом, выпавшим на верхней грани кубика. Все возможные события в данном случае перечислить сложно. Приведем некоторые примеры: выпадение тройки; выпадение четного числа; выпадение числа больше четырех.

Пример 2. Экспериментом является покупка и проверка лотерейного билета. Пусть часть билетов содержит выигрыш в 100 рублей, еще некоторое количество – приз в 1000 рублей, а по одному билету можно выиграть смартфон. Примеры событий: выигрыш приза в 100 рублей; выигрыш денежного приза; выигрыш смартфона; выигрыш; проигрыш.

Множество самых простых равновозможных событий, одно из которых обязательно произойдет при проведении эксперимента, называется ***множеством элементарных исходов*** и обозначается Ω, а каждое из таких событий называется ***элементарным исходом***.

Событием можно считать любое подмножество множества элементарных исходов. Таким образом, существует два подхода к определению понятия события. С одной стороны, событие – это явление, результат эксперимента, с другой – это множество, состоящее из элементарных исходов.

Выделим два особых события: $\varnothing$ – ***невозможное*** (событие, которое никогда не произойдет в условиях данного испытания); Ω – ***достоверное*** (событие, которое обязательно произойдет при испытании).

Продемонстрируем связь между двумя упомянутыми подходами к формализации понятия события на примерах.

Пример 1 (продолжение). Рассмотрим множество элементарных исходов и примеры событий при бросании кубика.

$\Omega = \{1;2;3;4;5;6\}$; $A = \{3\}$ – выпадение тройки; $B = \{2;4;6\}$ – выпадение четного числа; $C = \{5;6\}$ – выпадение числа больше 4.

Пример 2 (продолжение). В случае с лотереей множеством элементарных исходов является множество всех билетов. Событие «выигрыш 100 рублей» содержит те и только те билеты, выигрыш по которым составляет 100 рублей; событие «выигрыш» содержит выигрышные билеты, независимо от величины приза.

Пусть при некотором испытании в случае, если произошло событие A, то можно считать, что произошло также событие B. Тогда говорят, что A ***влечет за собой*** B (***из*** A ***следует*** B) или $A \subseteq B$. В этом случае множество A является подмножеством множества B. Убедимся в этом на примере.

Пример 1 (продолжение). Бросается кубик. Рассмотрим события: A – выпадение числа, не превосходящего 2; B – выпадение числа, не превосходящего 3. Очевидно, что A влечет за собой B.

При этом: $A = \{1;2\}$, $B = \{1;2;3\}$, откуда $A \subseteq B$.

1.2. Вероятность

В повседневной жизни под вероятностью понимают оценку шансов наступления события. В математике существует несколько подходов к формализации понятия вероятности. Рассмотрим некоторые из них.

Статистическая вероятность

Предположим, что существует возможность неоднократно воспроизвести эксперимент, и число произведенных испытаний равно N. Пусть требуемое событие A произошло в M случаях. Тогда под ***вероятностью*** события A будем понимать величину

$$P(A) = \lim_{N \to \infty} \frac{M}{N}.$$

Данное определение в целом соответствует интуитивному пониманию вероятности, однако сложно применимо на практике.

Классическая вероятность

Пусть множество элементарных исходов содержит N элементов (то есть, $|\Omega| = N$), а событие состоит из M элементов ($|A| = M \le N$). Исходы, составляющие A, называются ***благоприятными*** по отношению к этому событию. Тогда ***вероятностью*** события A будем называть величину

$$P(A) = \frac{M}{N}.$$

То есть, в классическом понимании вероятность – это отношение числа благоприятных элементарных исходов к общему числу исходов.

Пример 1 (продолжение). Найдем вероятность выпадения числа больше четырех при броске кубика.

$\Omega = \{1;2;3;4;5;6\}$, общее число исходов $N = 6$.

$C = \{5;6\}$, откуда число благоприятных исходов $M = 2$.

Следовательно, вероятность события C равна $P(C) = \frac{2}{6} = \frac{1}{3}$.

Из классического определения вытекают ***свойства вероятности***:

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

2) $P(\varnothing) = 0$ (вероятность невозможного события равна нулю);

3) $P(\Omega) = 1$ (вероятность достоверного события равна единице).

Геометрическая вероятность

Пусть множество элементарных исходов Ω бесконечно и представляет собой промежуток на числовой оси длины L. Событие A – промежуток или объединение промежутков суммарной длины l. Тогда ***вероятностью*** события A называется величина

$$P(A) = \frac{l}{L}.$$

Аналогично можно рассмотреть ситуацию, когда Ω и A – фигуры на плоскости, площади, соответственно, S_Ω и S_A. Тогда

$$P(A)=\frac{S_A}{S_\Omega}.$$

Пример 3. Автобусы ходят каждые пятнадцать минут. Какова вероятность того, что, подойдя к остановке наудачу, человек сможет уехать в ближайшие пять минут? Какова вероятность того, что он не сможет сесть в автобус в течение восьми минут?

Возможное время ожидания автобуса составляет от 0 до 15 минут от момента прихода на остановку, таким образом, $\Omega=[0;15]$ и $L=15$.

Исходы, при которых удастся уехать в течение пяти минут, лежат в интервале $A=[0;5]$ и $l_A=5$. Откуда $P(A)=\frac{5}{15}=\frac{1}{3}$.

Ситуация, когда уехать в течение восьми минут не удастся, означает, что автобус подойдет в интервале $B=[8;15]$. Следовательно, $l_B=7$ и $P(B)=\frac{7}{15}$.

1.3. Элементы комбинаторики

В сложных случаях для вычисления количества элементарных исходов необходимо использовать формулы комбинаторики.

Пусть элемент x можно выбрать из множества A n способами, а элемент y – из множества B m способами. Тогда верны следующие утверждения:

1) количество способов выбрать один элемент x или y равно $n+m$ – ***правило суммы***.

2) количество способов выбрать упорядоченную пару (x,y) (то есть x и y) равно $n\cdot m$ – ***правило произведения***.

Пусть множество A состоит из n элементов. ***Перестановкой из n элементов*** называется любая упорядоченная последовательность всех элементов множества A.

$P_n=n\cdot(n-1)\cdot(n-2)\cdot\ldots\cdot 2\cdot 1=n!$ – ***число перестановок из n элементов.***

Размещением из n по m (где $m\le n$) называется любая упорядоченная последовательность из m различных элементов множества A.

$$A_n^m = \frac{n!}{(n-m)!}$$ – ***число размещений из n по m***.

Сочетанием из n по m называется любой неупорядоченный набор из m различных элементов множества A.

$$C_n^m = \frac{n!}{m!(n-m)!}$$ – ***число сочетаний из n по m***.

В некоторых случаях необходимо рассмотреть размещения и сочетания, в которых элемент после появления может быть выбран повторно (например, грань кубика).

$$\widehat{A_n^m} = \underbrace{n \cdot n \cdot \ldots \cdot n}_{m \text{ раз}} = n^m$$ – ***число размещений с повторениями n по m***.

Пример 4. К экзамену нужно выучить материал по двенадцати темам. Студент освоил только восемь тем. В экзаменационном билете содержится два вопроса по двум различным темам. Какова вероятность того, что студент сможет ответить на оба вопроса? Хотя бы на один?

Элементарным исходом в данном случае является пара различных вопросов. Порядок, в котором вопросы располагаются в билете, неважен. Таким образом, билет представляет собой сочетание без повторения.

Общее число исходов находится как

$$N = C_{12}^2 = \frac{12!}{2! \cdot 10!} = \frac{12 \cdot 11}{2} = 66.$$

Число благоприятных исходов в первом случае – число билетов, составленных из вопросов только по тем темам, которые выучил студент:

$$M_A = C_8^2 = \frac{8!}{2! \cdot 6!} = \frac{8 \cdot 7}{2} = 28.$$

В итоге в первом случае вероятность равна

$$P(A) = \frac{28}{66} = \frac{14}{33} \approx 0,424.$$

Во втором случае необходимо рассмотреть ситуацию, когда студент знает ответ хотя бы на один вопрос в билете, то есть либо может ответить на оба вопроса, либо ответ на один ему известен, а на второй нет.

Количество билетов, в которых студент знает ответы на оба вопроса, уже найдено и равно $M_2 = M_A = 28$. Найдем количество билетов, в которых студент может ответить на один вопрос, а ответ на второй не выучил (таких вопросов четыре). В соответствии с правилом произведения оно равно

$$M_1 = C_8^1 \cdot C_4^1 = 8 \cdot 4 = 32.$$

Общее число билетов, удовлетворяющих описанию второго события, находится в соответствии с правилом суммы:

$$M_B = 28 + 32 = 60.$$

В итоге во втором случае вероятность равна

$$P(B) = \frac{60}{66} = \frac{10}{11} \approx 0{,}909.$$

Пример 5. Брошено три кубика. Какова вероятность того, что выпадет: **а)** три единицы; **б)** две единицы и двойка; **в)** единица, двойка и тройка?

Во всех трех случаях условия эксперимента одинаковы – бросается три кубика (или же один кубик бросается три раза – на дальнейшие рассуждения это не влияет). Элементарным исходом в данном случае является последовательность из трех чисел от 1 до 6 – размещение с повторением. Общее число исходов:

$$N = \widehat{A_6^3} = 6^3 = 216.$$

а) Существует только один благоприятный к описанному событию исход: (1,1,1). В итоге:

$$M_A = 1 \text{ и } P(A) = \frac{1}{216}.$$

б) Существует три благоприятных исхода: (1,1,2), (1,2,1), (2,1,1).

$$M_B = 3 \text{ и } P(B) = \frac{3}{216} = \frac{1}{72}.$$

в) Количество благоприятных исходов в данном случае равно числу перестановок из трех элементов.

$$M_C = 3! = 3 \cdot 2 = 6 \text{ и } P(C) = \frac{6}{216} = \frac{1}{36}.$$

В примере 5 вычисляются вероятности реализации различных сочетаний с повторениями при броске нескольких кубиков. Пример показывает, что эти вероятности не равны между собой. Таким образом, сочетания с повторением не являются элементарными исходами и для решения задач с использованием классического определения вероятности применены быть не могут. По этой причине формула числа сочетаний с повторениями в данном пособии не приводится.

1.4. Операции над событиями

Объединением событий A и B называется событие $A \cup B$, состоящее в том, что произошло хотя бы одно из событий, A или B.

Пересечением событий A и B называется событие $A \cap B$, состоящее в том, что произошли оба события, A и B.

Напомним, что объединение содержит те элементарные исходы, которые содержатся хотя бы в одном из множеств A или B, а пересечение содержит элементарные исходы, содержащиеся в A и B одновременно.

Объединение и пересечение называются также ***суммой*** ($A + B$) и ***произведением*** ($A \cdot B$) событий, соответственно [2, 3].

События A и B, которые не могут произойти одновременно (то есть, $A \cap B = \varnothing$), называются ***несовместными***. Если $A \cap B \neq \varnothing$, то события A и B называют ***совместными***.

Два несовместных события, одно из которых обязательно произойдет, называются ***противоположными***. Событие, противоположное к A, обозначается $\overline{A}$. Иногда говорят, что противоположное событие $\overline{A}$ состоит в том, что A при испытании не произошло.

Можно также сказать, что противоположным является событие, обладающее свойствами:

1) $\overline{A} \cap A = \varnothing$; **2)** $\overline{A} \cup A = \Omega$; **3)** $\overline{\left(\overline{A}\right)} = A$.

Пример 6. Бросается кубик. Рассмотрим событие A и противоположное к нему: $A = \{3;6\}$ – выпадение числа, делящегося на три; $\overline{A} = \{1;2;4;5\}$ – выпадение числа, не делящегося на три.

Далее будут представлены теоремы, позволяющие вычислять вероятности объединения, пересечения и противоположного события.

1.5. Теоремы сложения вероятностей

Теорема 1 (теорема сложения вероятностей для несовместных событий). Если события A и B несовместны (то есть, $A \cap B = \varnothing$), то

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

Теорема 2 (теорема сложения вероятностей для совместных событий). Если события A и B совместны (то есть, $A \cap B \neq \varnothing$), то

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

Формула для вычисления вероятности объединения большого числа совместных событий в данном курсе рассматриваться не будет. Если возникает необходимость ее использования, рекомендуем применять теорему о вероятности противоположного события.

Теорема 3 (вероятность противоположного события). Для любого события A верно

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

Пример 7. Бросается четыре монетки. Найти вероятность того, что выпадет хотя бы один орел.

В данной задаче необходимо вычислить вероятность объединения четырех совместных событий (выпадение орла на первой монетке, выпадение орла на второй и т. д.). Рассмотрим событие D, противоположное к заданному в условии. Это событие состоит в том, что при броске четырех монеток выпадет четыре решки.

Общее число элементарных исходов при броске четырех монеток $N = \widehat{A_2^4} = 2^4 = 16$. Благоприятный исход один – четыре решки, $M = 1$. В итоге $P(D) = \dfrac{1}{16}$.

В соответствии с теоремой 3, $P(\overline{D}) = 1 - P(D) = 1 - \dfrac{1}{16} = \dfrac{15}{16}$.

1.6. Теоремы умножения вероятностей

Условной вероятностью $P(A|B)$ события A при условии события B называется вероятность A в случае, если известно, что B произошло.

Пример 8. Бросается кубик. Рассмотрим следующие события: $A=\{3\}$ – выпадение тройки; $B=\{1;3;5\}$ – выпадение нечетного числа; $C=\{5;6\}$ – выпадение числа больше четырех.

Очевидно, что $P(A)=\frac{1}{6}$, $P(B)=\frac{3}{6}=\frac{1}{2}$, $P(C)=\frac{2}{6}=\frac{1}{3}$.

Тогда $P(A|B)=\frac{1}{3}$ (выпадение тройки, если известно, что выпало нечетное число); $P(B|A)=1$ (выпадение нечетного числа при условии, что выпала тройка); $P(A|C)=0$ (выпадение тройки при условии, что выпало число больше четырех); $P(C|A)=0$ (выпадение числа больше четырех при условии, что выпала тройка).

Теорема 4 (теорема умножения вероятностей). Для любых событий A и B верно

$$P(A\cap B)=P(A)\cdot P(B|A)=P(B)\cdot P(A|B).$$

События называются ***независимыми***, если в случае, когда произошло одно из них, вероятность другого не меняется, то есть если $P(A|B)=P(A)$ и $P(B|A)=P(B)$.

Независимыми являются, например, события, описывающие результаты бросков двух кубиков (или повторных бросков одного кубика), а также выигрыши в двух разных лотереях.

Теорема 5 (теорема умножения вероятностей для независимых событий). Если события A и B независимы, то

$$P(A\cap B)=P(A)\cdot P(B).$$

Пример 9. Два школьника тренируются забрасывать мяч в баскетбольную корзину. Известно, что первый школьник попадает в корзину в одном случае из трех, второй – в половине случаев. Школьники сделали по одному броску. Какова вероятность того, что удачным был хотя бы один.

Рассмотрим следующие события: A – первый школьник попал в корзину; B – второй школьник попал в корзину.

В задаче требуется найти вероятность объединения событий A и B. Эти события являются совместными, и для решения необходимо использовать теорему 2.

Из условия задачи следует, что $P(A)=\frac{1}{3}$, $P(B)=\frac{1}{2}$.

Так как описанные в задаче события независимы, то вероятность их пересечения вычисляется по теореме 5:

$$P(A\cap B)=P(A)\cdot P(B)=\frac{1}{3}\cdot\frac{1}{2}=\frac{1}{6}.$$

В соответствии с теоремой 2, вероятность того, что хотя бы один бросок окажется удачным, будет следующей:

$$P(A\cup B)=P(A)+P(B)-P(A\cap B)=\frac{1}{3}+\frac{1}{2}-\frac{1}{6}=\frac{2+3-1}{6}=\frac{4}{6}=\frac{2}{3}.$$

Существует еще один способ решения данной задачи. Рассмотрим событие, противоположное описанному в условии, – оба броска были неудачными. Найдем его вероятность:

$$P(D)=P(\overline{A}\cap\overline{B})=\left|\begin{array}{l}\text{по теореме 5 так как}\\ A \text{ и } B \text{ независимы}\end{array}\right|=P(\overline{A})\cdot P(\overline{B})=\frac{2}{3}\cdot\frac{1}{2}=\frac{1}{3}.$$

Следовательно, искомая вероятность равна

$$P(\overline{D})=\left|\text{по теореме 3}\right|=1-P(D)=1-\frac{1}{3}=\frac{2}{3}.$$

Теорема 5 верна также для пересечения трех и более независимых событий. То есть, если события $A_1,A_2,...,A_n$ попарно независимы, то $P\left(\bigcap_{k=1}^{n}A_k\right)=\prod_{k=1}^{n}P(A_k)$.

1.7. Формула полной вероятности. Формула Байеса

Говорят, что события $H_1,H_2,...,H_n$ образуют ***полную группу***, если они попарно несовместны и их объединение равно множеству элементарных исходов. Иными словами, при некотором испытании происходит одно и только одно из событий, образующих полную группу.

Если $H_1,H_2,...,H_n$ – полная группа, то $\sum_{i=1}^{n}P(H_i)=1$.

Теорема 6 (формула полной вероятности). Если событие A происходит при условии того, что случилось одно из событий $H_1, H_2, ..., H_n$, образующих полную группу, то вероятность этого события может быть найдена по формуле:

$$P(A) = \sum_{i=1}^{n} P(H_i) \cdot P(A \mid H_i).$$

Пример 10. На завод поступило две партии однотипных деталей. Объем первой партии составляет 1 000 деталей, второй – 1 500 деталей. При этом первая партия содержит 5 % брака, вторая 8 %. Какова вероятность того, что взятая наудачу деталь окажется качественной?

Рассмотрим следующие события:

A – взятая наудачу деталь является качественной;

H_1 – взятая наудачу деталь – из первой партии;

H_2 – взятая наудачу деталь – из второй партии.

События H_1 и H_2 образуют полную группу. Найдем их вероятности в соответствии с классическим определением:

$$P(H_1) = \frac{1000}{1000+1500} = \frac{1000}{2500} = 0{,}4, \; P(H_2) = \frac{1500}{1000+1500} = \frac{1500}{2500} = 0{,}6.$$

Условная вероятность события A при условии H_1 составляет 0,95, а условная вероятность A при условии H_2 составляет 0,92. То есть

$$P(A \mid H_1) = 0{,}95; \; P(A \mid H_2) = 0{,}92.$$

В соответствии с формулой полной вероятности

$$P(A) = P(H_1) \cdot P(A \mid H_1) + P(H_2) \cdot P(A \mid H_2) = 0{,}4 \cdot 0{,}95 + 0{,}6 \cdot 0{,}92 = 0{,}932.$$

Пусть требуется найти вероятность одного из событий полной группы при условии, что событие A произошло. Для решения этой задачи используется ***формула Байеса***:

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

Пример 10 (продолжение). Взятая наудачу деталь оказалась качественной. Какова вероятность, что она из первой партии? Взятая наудачу деталь бракованная. Какова вероятность, что она из второй партии?

Чтобы ответить на первый вопрос, применим формулу Байеса:

$$P\left(H_1 \mid A\right)=\frac{P\left(H_1\right)\cdot P\left(A \mid H_1\right)}{P\left(A\right)}=\frac{0,4\cdot 0,95}{0,932}\approx 0,408\,.$$

Применим ту же формулу для ответа на второй вопрос:

$$P\left(H_2 \mid \overline{A}\right)=\frac{P\left(H_2\right)\cdot P\left(\overline{A} \mid H_2\right)}{P\left(\overline{A}\right)}=\frac{0,6\cdot 0,08}{0,068}\approx 0,706\,,$$

где $P\left(\overline{A}\right)=1-P\left(A\right)=1-0,932=0,068$ – вероятность того, что взятая наудачу деталь бракованная; $P\left(\overline{A} \mid H_2\right)=0,08$ – вероятность того, что деталь бракованная при условии, что она из второй партии.

2. СЛУЧАЙНЫЕ ВЕЛИЧИНЫ

2.1. Понятие случайной величины. Дискретные случайные величины

Случайной величиной называется переменная, которая в результате испытания принимает одно из множества возможных числовых значений, предсказать которое заранее невозможно.

Дискретной называется случайная величина, которая принимает конечное или счетное число значений. ***Непрерывной*** – случайная величина, значения которой образуют интервал на числовой оси.

Пример 11. Примерами случайных величин являются: **а)** результат броска кубика; **б)** количество бракованных изделий в партии товара (шт.); **в)** рост взятого наудачу человека (см). Дискретными являются случайные величины, перечисленные в пунктах **а)** и **б)**; непрерывной – из пункта **в)**.

Случайные величины обычно обозначаются заглавными латинскими буквами из конца алфавита, например, X, Y, Z, а принимаемые ими значения – соответствующими строчными буквами: x, y, z, иногда с использованием индексов.

Каждая случайная величина задается некоторым ***законом распределения*** – описанием, устанавливающим связь между значениями случайной величины и соответствующими им вероятностями.

Для дискретной случайной величины закон распределения может выглядеть как таблица, формула или график. Таблица, называемая также ***рядом распределения***, имеет вид:

X:	x_i	x_1	x_2	…	x_n
	p_i	p_1	p_2	…	p_n

События $X = x_i$ при $i = 1, 2, ..., n$ образуют полную группу, отчего в соответствии с теоремой 1 верно $\sum_{i=1}^{n} p_i = 1$. Данное равенство называется ***основным свойством ряда распределения***.

Ряд распределения может быть представлен на графике в виде ломаной, называемой ***полигоном распределения вероятностей*** (рис. 1).

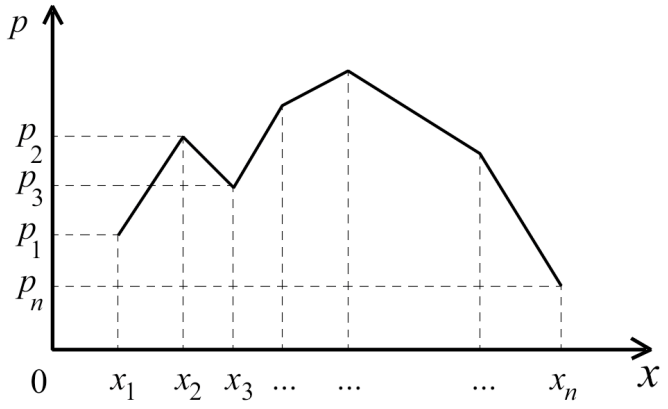


Рис. 1. Полигон распределения вероятностей

Пример 12. Вероятность заключения договора на установку окон после визита замерщика к потенциальному клиенту составляет 0,7. Составить закон распределения числа заключенных договоров после выезда замерщиком по двум адресам. Представить закон распределения в виде таблицы и полигона распределения вероятностей.

Пусть X – число заключенных договоров. Эта случайная величина может принимать значения 0, 1, 2. Для вычисления их вероятностей рассмотрим события: A_1 – первый из возможный клиентов заключил договор; A_2 – второй из возможный клиентов заключил договор.

$$P(X=0) = P\left(\overline{A_1} \cap \overline{A_2}\right) = \left|\text{по теор. 5}\right| = P\left(\overline{A_1}\right) \cdot P\left(\overline{A_2}\right) = 0,3 \cdot 0,3 = 0,09;$$

$$P(X=1) = P\left(\left(A_1 \cap \overline{A_2}\right) \cup \left(\overline{A_1} \cap A_2\right)\right) = \left|\text{по теор. 1}\right| = P\left(\left(A_1 \cap \overline{A_2}\right)\right) +$$

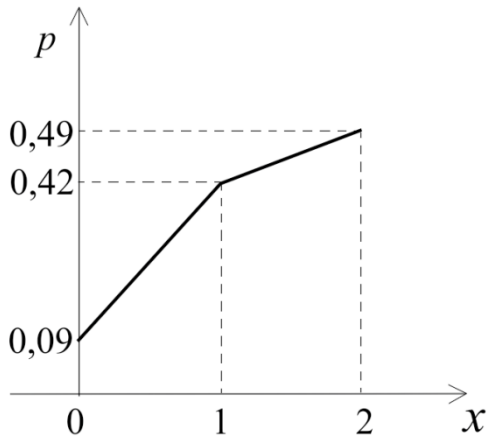
$$+P\left(\left(\overline{A_1} \cap A_2\right)\right) = P(A_1) \cdot P\left(\overline{A_2}\right) + P\left(\overline{A_1}\right) \cdot P(A_2) = 0,7 \cdot 0,3 + 0,3 \cdot 0,7 = 0,42;$$

$$P(X=2) = P(A_1 \cap A_2) = P(A_1) \cdot P(A_2) = 0,7 \cdot 0,7 = 0,49.$$

В итоге таблица и полигон распределения вероятностей выглядят так:

X :

x_i	0	1	2
p_i	0,09	0,42	0,49



Две случайных величины называются ***независимым***, если значения, которая принимает каждая из них, не влияют на закон распределения другой. Примерами независимых случайных величин могут служить выигрыши в двух различных лотереях; результаты двух бросков кубика.

В противном случае случайные величины называются ***зависимыми***. Примеры зависимых случайных величин: рост наудачу взятого человека и его вес; количество электроприборов в квартире и расход электроэнергии жильцами квартиры в течение месяца.

Пусть даны случайные величины X и Y такие, что

X :

x_1	x_2	…	x_n
p_1	p_2	…	p_n

Y :

y_1	y_2	…	y_m
q_1	q_2	…	q_m

Произведением случайной величины X ***на число*** c называется случайная величина cX, распределенная в соответствии с законом:

cX :

cx_1	cx_2	…	cx_n
p_1	p_2	…	p_n

k-й степенью случайной величины X называется случайная величина X^k, распределенная в соответствии с законом:

X^k :

x_1^k	x_2^k	…	x_n^k
p_1	p_2	…	p_n

Суммой (или ***произведением***) случайных величин X и Y называется случайная величина $X+Y$ ($X \cdot Y$), которая принимает все возможные

значения $x_i + y_j$ $(x_i \cdot y_j)$ с вероятностями $p_{ij} = P\left(\{X = x_i\} \cap \{Y = y_j\}\right)$, где $i = 1,...,n$, $j = 1,...,m$.

Если случайные величины X и Y независимы, то независимы и события $\{X = x_i\}$ и $\{Y = y_j\}$. Тогда в силу теоремы 5, $p_{ij} = P(X = x_i) \cdot P(Y = y_j) = p_i \cdot q_j$.

Пример 13. Даны случайные величины X и Y, заданные рядами распределения:

X:

x_i	1	2
p_i	0,4	0,6

Y:

y_i	– 2	1	2
q_i	0,5	0,3	0,2

Найти законы распределения следующих случайных величин:

а) $U = 3X$; **б)** $V = Y^2$; **в)** $W = X + Y$.

а) Каждое из значений случайной величины X умножается на 3. В итоге:

U:

u_i	3	6
p_i	0,4	0,6

б) В данном случае случайная величина V может принимать значение 4 в двух случаях – если значение $Y = 2$ или $Y = -2$. Данные события являются несовместными. В соответствии с теоремой 1:

$P(V = 4) = P\left(\{Y = 2\} \cup \{Y = -2\}\right) =$

$= P(Y = 2) + P(Y = -2) = 0,5 + 0,2 = 0,7$ и

V:

v_i	1	4
p_i	0,3	0,7

в) Для нахождения закона распределения суммы двух случайных величин построим таблицу, в которой перечислены все возможные пары значений x_i и y_j, для каждой пары вычислим сумму и вероятность ее появления в соответствии с теоремой 5.

X	Y	$W = X + Y$	P
1	– 2	– 1	0,4·0,5 = 0,2
1	1	2	0,4·0,3 = 0,12
1	2	3	0,4·0,2 = 0,08
2	– 2	0	0,6·0,5 = 0,3
2	1	3	0,6·0,3 = 0,18
2	2	4	0,6·0,2 = 0,12

Таким образом, случайная величина W принимает значения –1, 0, 2, 3, 4. Заметим, что значение 3 возможно в двух случаях – для пары (1, 2) и для пары (2, 1). По теореме 1:

$$P(W=3)=P\big((1,2)\cup(2,1)\big)=P(1,2)+P(2,1)=0,08+0,18=0,26.$$

Закон распределения для W выглядит следующим образом:

W :

w_i	– 1	0	2	3	4
p_i	0,2	0,3	0,12	0,26	0,12

2.2. Числовые характеристики дискретной случайной величины

Таблица (ряд распределения) дает полное, исчерпывающее описание дискретной случайной величины. Однако для того, чтобы изучить поведение случайной величины в случае многократной ее реализации, требуется дополнительный анализ ряда распределения.

Рассмотрим следующую задачу.

Пример 14. Пусть на соревнованиях по стрельбе из лука результаты одного выстрела оцениваются следующим образом: за попадание по мишени дается 1 балл, за попадание в центр мишени («яблочко») начисляется 3 балла, промах оценивается в 0 баллов. Наблюдения за двумя стрелками показали, что результаты их выстрелов описываются случайными величинами X и Y со следующими рядами распределения.

X :

x_i	0	1	3
p_i	0,2	0,5	0,3

Y :

y_i	0	1	3
p_i	0,3	0,35	0,35

Какой из стрелков стреляет лучше?

Для того, чтобы сделать заключение, необходимо найти среднее количество баллов, которое получает каждый из стрелков за один выстрел.

Математическим ожиданием (***средним значением***) случайной величины X называется число MX, равное сумме произведений всех ее значений на их вероятности, то есть:

$$MX=\sum_{i=1}^{n}x_i p_i .$$

Встречаются также следующие обозначения: $M(X)$, EX, $E(X)$.

Пример 14 (продолжение). Найдем математические ожидания случайных величин X и Y.

$$MX = 0 \cdot 0,2 + 1 \cdot 0,5 + 3 \cdot 0,3 = 0,5 + 0,9 = 1,4;$$
$$MY = 0 \cdot 0,3 + 1 \cdot 0,35 + 3 \cdot 0,35 = 4 \cdot 0,35 = 1,4.$$

Среднее число баллов, получаемое за один выстрел, у обоих стрелков одинаковое. В среднем оба спортсмена стреляют одинаково хорошо.

Перечислим ***свойства математического ожидания***.

1) $MC = C$, где $C = const$ – математическое ожидание константы равно этой константе (константу можно рассматривать как случайную величину, принимающую единственное значение C с вероятностью 1).

2) $M(CX) = C \cdot MX$, где $C = \text{const}$ – числовой множитель можно выносить за знак математического ожидания.

3) $M(X \pm Y) = MX \pm MY$ – математическое ожидание суммы (разности) равно сумме (разности) математических ожиданий.

4) $M(XY) = MX \cdot MY$, если X и Y независимы – математическое ожидание произведения независимых случайных величин равно произведению их математических ожиданий.

5) $M(X \pm C) = MX \pm C$ – если значение случайной величины изменить на константу, то ее математическое ожидание соответствующим образом изменится на ту же константу.

6) $M(X - MX) = 0$. Случайная величина $X - MX$ описывает отклонение случайной величины X от ее среднего значения. Она положительна, если исходная случайная величина принимает значение больше среднего, и отрицательна, если это значение меньше среднего. В итоге в среднем значение отклонения равно нулю.

И все же иногда требуется исследовать именно отклонение случайной величины X от ее среднего, чтобы оценить разброс принимаемых ею значений. Численной характеристикой разброса значений случайной величины является дисперсия.

Дисперсией случайной величины X называется математическое ожидание квадрата ее отклонения от среднего, то есть:

$$DX = M\left(X - MX\right)^2.$$

Дисперсия описывает квадрат отклонения, что создает определенные неудобства. Поэтому в качестве показателя разброса случайной величины используется квадратный корень из дисперсии или ***среднее квадратическое отклонение*** (***стандартное отклонение***):

$$\sigma_X = \sqrt{DX}.$$

Дисперсия дискретной случайной величины вычисляется по формуле:

$$DX = \sum_{i=1}^{n}\left(x_i - MX\right)^2 p_i.$$

Пример 14 (продолжение). Найти дисперсию случайной величины X.

$$DX = \left(0-1,4\right)^2 \cdot 0,2 + \left(1-1,4\right)^2 \cdot 0,5 + \left(3-1,4\right)^2 \cdot 0,3 = 1,34.$$

Перечислим ***<u>свойства дисперсии</u>***.

1) $DC = 0$, где $C = \text{const}$ – дисперсия константы равна нулю.

2) $D\left(CX\right) = C^2 \cdot DX$, где $C = \text{const}$ – числовой множитель можно выносить за знак дисперсии, возведя его в квадрат.

3) $D\left(X \pm Y\right) = DX + DY$, если X и Y независимы – дисперсия суммы (разности) независимых случайных величин равна сумме их дисперсий.

4) $DX = MX^2 - \left(MX\right)^2$.

Величина MX^k называется ***k-м начальным моментом случайной величины*** X. Для вычисления дисперсии используется второй момент. Величина $M\left(X - MX\right)^k$ носит название ***k-го центрального момента***. Таким образом, дисперсия – это второй центральный момент.

Пример 14 (продолжение). Найти дисперсию Y.

По свойству 4, $DY = MY^2 - \left(MY\right)^2$. В соответствии с п. 2.1., чтобы найти квадрат случайной величины, необходимо ее значения возвести в квадрат, а их вероятности остаются неизменным. Тогда

$$MY^2 = 0^2 \cdot 0,3 + 1^2 \cdot 0,35 + 3^2 \cdot 0,35 = 10 \cdot 0,35 = 3,5.$$

Вычислим дисперсию:

$$DY = 3,5 - (1,4)^2 = 3,5 - 1,96 = 1,54.$$

В итоге дисперсия второго стрелка из лука больше, чем у первого. Это означает, что результаты первого спортсмена более стабильны, и позволяет заключить, что он стреляет лучше.

Модой случайной величины (MoX) называется ее наиболее вероятное значение. Для дискретной случайной величины X мода – это значение x_i, которому соответствует наибольшая вероятность p_i .

Пример 14 (продолжение). Для случайной величины X максимальной является вероятность $P(1) = 0,5$, откуда $MoX = 1$.

Для случайной величины Y существует два значения, 1 и 3, вероятности которых равны между собой и максимальны. В итоге случайная величина имеет две моды: $Mo_1Y = 1$ и $Mo_2Y = 3$.

2.3. Функция распределения

Функцией распределения случайной величины X называется функция, определенная на множестве действительных чисел и выражающая вероятность того, что случайного величина примет значение меньшее, чем x:

$$F(x) = P(X < x).$$

Пример 12 (продолжение). Построить функцию распределения случайной величины, описанной в примере 12. п. 2.1.

Закон распределения данной случайной величины задан таблицей:

X :

x_i	0	1	2
p_i	0,09	0,42	0,49

Поведение функции распределения меняется в точках 0, 1 и 2, которые разбивают числовую ось на четыре промежутка.

Рассмотрим функцию на каждом из них.

1) $x \le 0$. Возьмем любую точку промежутка, например, $x = -1$, и вычислим значение функции распределения в этой точке в соответствии с

определением. $F(-1)=P(X<-1)=0$ – так случайная величина не имеет значений меньше –1. Нулевое значение функция распределения будет иметь в любой точке данного промежутка.

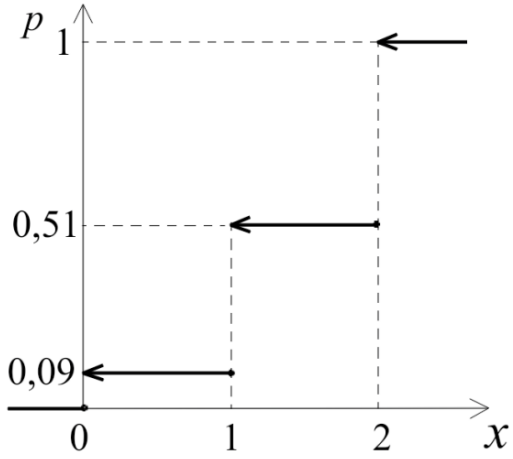
2) $0<x\le 1$. Возьмем любую точку промежутка, например $x=0,4$. Для нее, а также для каждой точки промежутка, включая 1, $F(0,4)=P(X<0,4)=P(X=0)=0,09$.

3) $1<x\le 2$, например $x=2$. В соответствии с теоремой 1, значение функции: $F(2)=P(X<2)=P(\{X=0\}\cup\{X=1\})=0,09+0,42=0,51$.

4) $x>2$, например $x=4$. Здесь $F(4)=P(X<4)=1$, так как все значение случайной величины меньше 4.

Функция распределения и ее график будут следующими:

$$F(x)=\begin{cases}0, & \text{при} \quad x\le 0;\\ 0,09, & \text{при} \quad 0<x\le 1;\\ 0,51, & \text{при} \quad 1<x\le 2;\\ 1, & \text{при} \quad x>2.\end{cases}$$



Пример показывает, что для дискретной случайной величины функция распределения является кусочно-постоянной (ступенчатой) и непрерывной слева.

Перечислим ***<u>свойства функции распределения</u>***:

1) $0\le F(x)\le 1$ – верно, так как $F(x)$ – вероятность;

2) $F(x)$ – неубывающая функция;

3) $F(-\infty)=\lim\limits_{x\to-\infty}F(x)=P(X<-\infty)=0$,

$F(+\infty)=\lim\limits_{x\to+\infty}F(x)=P(X<+\infty)=1$;

4) $P(x_1\le X<x_2)=F(x_2)-F(x_1)$.

Пример 12 (продолжение). Найти вероятность того, что случайная величина примет значение из промежутка $[0,2;\ 2,7)$.

По свойству 4 функции распределения имеем:

$$P(0,2 \le X < 2,7) = F(2,7) - F(0,2) = 1 - 0,09 = 0,91.$$

Можно решить ту же задачу иначе. В соответствии с теоремой 1:

$$P(0,2 \le X < 2,7) = P(1) + P(2) = 0,42 + 0,49 = 0,91.$$

2.4. Непрерывные случайные величины

Из п. 2.1. известно, что непрерывной называется случайная величина, значения которой образуют промежуток на числовой оси. Функция распределения непрерывной случайной величины является непрерывной на всем множестве действительных чисел и дифференцируемой во всех точках, кроме, может быть, отдельных точек. Такие точки на графике функции выглядят как точки излома.

Функция распределения случайной величины, принимающей значения из отрезка $[a;b]$, может выглядеть так, как изображено на рис. 2.

Теорема 7. Вероятность любого отдельно взятого значения непрерывной случайной величины равна нулю.

С учетом перечисленных особенностей уточним ***свойства непрерывной функции распределения***:

1) $0 \le F(x) \le 1$;

2) $F'(x) \ge 0$;

3) $F(-\infty) = 0$, $F(+\infty) = 1$;

4) $P(x_1 \le X \le x_2) = P(x_1 < X < x_2) =$

$= P(x_1 \le X < x_2) = P(x_1 < X \le x_2) = F(x_2) - F(x_1)$.

Закон распределения непрерывной случайной величины также может быть представлен ***плотностью распределения*** или ***плотностью вероятности*** – функцией $f(x)$, равной производной функции распределения случайной величины, $f(x) = F'(x)$.

Плотность распределения случайной величины, принимающей значения из отрезка $[a;b]$, может выглядеть так, как изображено на рис. 3. Функция $f(x)$ здесь отлична от нуля только на отрезке $[a;b]$. Значения x, которым соответствуют большие значения функции, являются более вероятными, чем значения, которым соответствуют меньшие значения $f(x)$. Таким образом, случайная величина, заданная плотностью, график которой изображен на рис. 3, чаще принимает значения, близкие к b, и реже близкие к a.

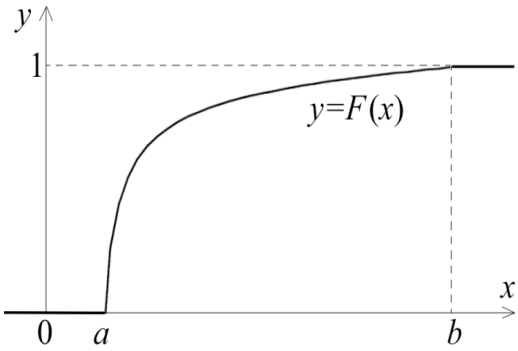


Рис. 2. Функция распределения

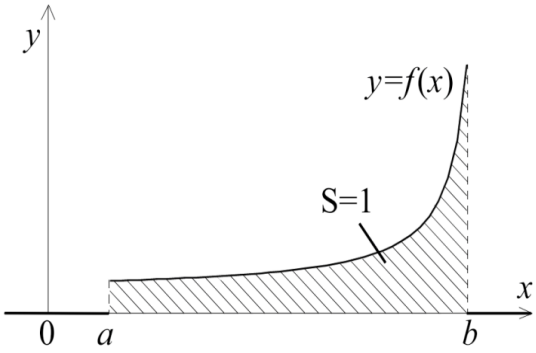


Рис. 3. Плотность распределения

Перечислим ***<u>свойства плотности распределения</u>***:

1) $f(x) \geq 0$ (свойство 2 непрерывной функции распределения);

2) $\int\limits_{-\infty}^{+\infty} f(x)dx = 1$ (рис. 3);

3) $F(x) = \int\limits_{-\infty}^{x} f(t)dt$ (рис. 4);

4) $P(x_1 \leq X \leq x_2) = P(x_1 < X < x_2) = P(x_1 \leq X < x_2) =$

$= P(x_1 < X \leq x_2) = \int\limits_{x_1}^{x_2} f(x)dx = F(x)\Big|_{x_1}^{x_2} = F(x_2) - F(x_1)$ (рис. 5).

Плотность распределения иногда называют дифференциальной функцией распределения, а саму функцию распределения – интегральной функцией распределения [3, 5].

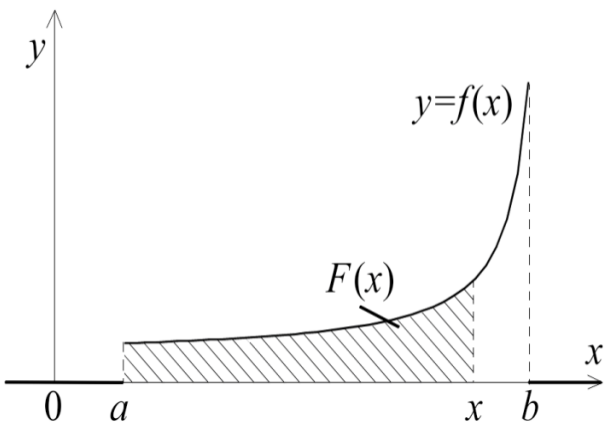


Рис. 4. Плотность распределения и функция распределения

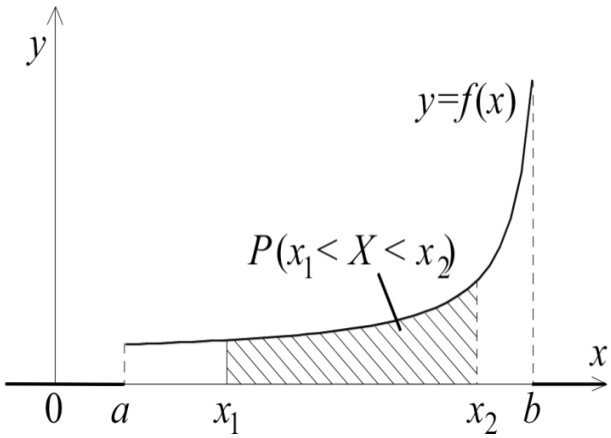


Рис. 5. Плотность распределения и вероятность

Пример 15. Определить, для какого значения параметра a функция

$$f(x)=\begin{cases} 0, & \text{при} \quad x<1; \\ \dfrac{a}{x^4}, & \text{при} \quad x\geq 1. \end{cases}$$

является плотностью распределения некоторой случайной величины. Найти функцию распределения этой случайной величины и вероятность того, что случайная величина примет значение из интервала $(2;4)$.

Для того, чтобы функция являлась плотностью распределения, необходимо выполнение свойств плотности 1 и 2.

Очевидно, что свойство 1 выполнено при $a>0$. Для проверки свойства 2 вычислим интеграл функции по всей числовой прямой. Он должен быть равен единице.

$$\int_{-\infty}^{+\infty} f(x)dx=\int_{-\infty}^{1} 0dx+\int_{1}^{+\infty}\frac{a}{x^4}dx=0+\lim_{\beta\to+\infty} a\int_{1}^{\beta} x^{-4}dx=\lim_{\beta\to+\infty} a\left.\frac{x^{-3}}{-3}\right|_{1}^{\beta}=$$

$$=\lim_{\beta\to+\infty}-\left.\frac{a}{3x^3}\right|_{1}^{\beta}=\lim_{\beta\to+\infty}\left(-\frac{a}{3\beta^3}+\frac{a}{3\cdot 1^3}\right)=0+\frac{a}{3}=1, \text{ откуда } a=3.$$

В итоге функция плотности примет вид:

$$f(x)=\begin{cases} 0, & \text{при} \quad x<1; \\ \dfrac{3}{x^4}, & \text{при} \quad x\ge 1. \end{cases}$$

Функцию распределения найдем по свойству 3. Рассмотрим ее в отдельности на каждом из на двух промежутков.

$$x<1:\ F(x)=\int_{-\infty}^{x} 0dt=0;$$

$$x\ge 1:\ F(x)=\int_{-\infty}^{1} 0dt+\int_{1}^{x}\frac{3}{t^4}dt=0+3\int_{1}^{x}t^{-4}dx=3\cdot\left.\frac{t^{-3}}{-3}\right|_{1}^{x}=-\left.\frac{1}{t^3}\right|_{1}^{x}=-\frac{1}{x^3}+1.$$

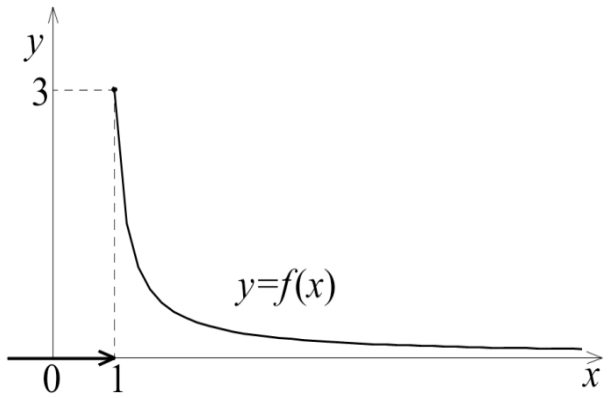
В итоге функция распределения выглядит следующим образом:

$$F(x)=\begin{cases} 0, & \text{при} \quad x<1; \\ 1-\dfrac{1}{x^3}, & \text{при} \quad x\ge 1. \end{cases}$$

Для вычисления вероятности попадания в интервал используем свойство 4 непрерывной функции распределения:

$$P(2<X\le 4)=F(4)-F(2)=\left(1-\frac{1}{4^3}\right)-\left(1-\frac{1}{2^3}\right)=\frac{1}{8}-\frac{1}{64}=\frac{7}{64}.$$

Графики функций $f(x)$ и $F(x)$ для заданной случайной величины будут выглядеть так:



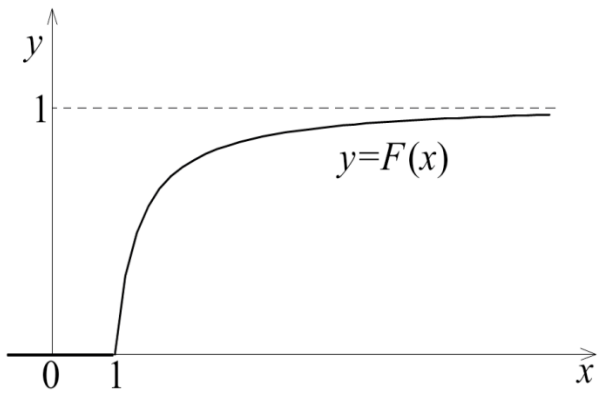


Рис. 6. Пример 15. Плотность распределения

Рис. 7. Пример 15. Функция распределения

2.5. Числовые характеристики непрерывной случайной величины

Математическое ожидание и дисперсия непрерывной случайной величины вычисляются в соответствии с формулами:

$$MX = \int_{-\infty}^{+\infty} x \cdot f(x)\,dx; \qquad DX = \int_{-\infty}^{+\infty} (x - MX)^2 \cdot f(x)\,dx.$$

Кроме того, дисперсия может быть найдена по формуле

$$DX = MX^2 - (MX)^2, \text{ где } MX^2 = \int_{-\infty}^{+\infty} x^2 \cdot f(x)\,dx.$$

Модой MoX (наиболее вероятным значением) непрерывной случайной величины является точка максимума плотности распределения.

Медианой MeX непрерывной случайной величины называется значение $MeX = x^*$, для которого $P\left(X > x^*\right) = P\left(X < x^*\right) = 0,5$. Иными словами, медиана – это точка, разбивающая интервал значений случайной величины на два промежутка, таких, что случайная величина с равной вероятностью принимает значение в каждом из них.

Так как по определению функции распределения $F(x) = P(X < x)$, то для медианы $MeX = x^*$ верно равенство $F\left(x^*\right) = 0,5$.

Пример 15 (продолжение). Вычислить математическое ожидание, дисперсию, среднее квадратическое отклонение, моду и медиану для случайной величины, заданной плотностью распределения

$$f(x) = \begin{cases} 0, & \text{при} \quad x < 1; \\ \dfrac{3}{x^4}, & \text{при} \quad x \geq 1. \end{cases}$$

Найдем математическое ожидание:

$$MX=\int_{-\infty}^{+\infty}x\cdot f(x)dx=\int_{-\infty}^{1}x\cdot 0dx+\int_{1}^{+\infty}x\cdot\frac{3}{x^4}dx=3\int_{1}^{+\infty}\frac{1}{x^3}dx=3\lim_{\beta\to+\infty}\int_{1}^{\beta}x^{-3}dx=$$

$$=3\lim_{\beta\to+\infty}\left.\frac{x^{-2}}{-2}\right|_1^{\beta}=3\lim_{\beta\to+\infty}\left.\left(-\frac{1}{2x^2}\right)\right|_1^{\beta}=3\lim_{\beta\to+\infty}\left(-\frac{1}{2\beta^2}+\frac{1}{2\cdot 1^2}\right)=3\cdot\frac{1}{2}=\frac{3}{2}.$$

Найдем дисперсию:

$$MX^2=\int_{-\infty}^{+\infty}x^2\cdot f(x)dx=\int_{-\infty}^{1}x^2\cdot 0dx+\int_{1}^{+\infty}x^2\cdot\frac{3}{x^4}dx=3\int_{1}^{+\infty}\frac{1}{x^2}dx=3\lim_{\beta\to+\infty}\int_{1}^{\beta}x^{-2}dx=$$

$$=3\lim_{\beta\to+\infty}\left.\frac{x^{-1}}{-1}\right|_1^{\beta}=3\lim_{\beta\to+\infty}\left.\left(-\frac{1}{x}\right)\right|_1^{\beta}=3\lim_{\beta\to+\infty}\left(-\frac{1}{\beta}+\frac{1}{1}\right)=3,\text{ откуда}$$

$$DX=MX^2-(MX)^2=3-\left(\frac{3}{2}\right)^2=3-\frac{9}{4}=\frac{12-9}{4}=\frac{3}{4}.$$

В соответствии с п. 2.2, среднее квадратическое отклонение равно

$$\sigma_X=\sqrt{DX}=\sqrt{\frac{3}{4}}=\frac{\sqrt{3}}{2}.$$

Из рис. 6 можно заключить, что максимум плотности распределения достигается в точке 1, которая и является модой,

$$MoX=1.$$

Для вычисления медианы рассмотрим функцию распределения:

$$F(x)=\begin{cases}0, & \text{при} \quad x<1;\\ 1-\dfrac{1}{x^3}, & \text{при} \quad x\ge 1.\end{cases}$$

Значение 0,5 эта функция может принимать при $x\ge 1$. Найдем медиану из уравнения $1-\frac{1}{x^3}=\frac{1}{2}$, откуда $\frac{1}{x^3}=\frac{1}{2}$, следовательно $x^3=2$, $x=\sqrt[3]{2}$ и $MeX=\sqrt[3]{2}$.

2.6. Основные законы распределения

Рассмотрим законы распределения, наиболее часто используемые для решения научных и практических задач. Законы распределения I–VI задают дискретные случайные величины, а VI–VIII – непрерывные.

I. Распределение Бернулли с параметром *p* (схема Бернулли) – B_p

Пусть проводится испытание, в котором некоторое событие происходит с вероятностью p, а с вероятностью $q = 1 - p$ испытание завершается неудачей. Пусть X – количество успехов в одном таком испытании. Тогда говорят, что случайная величина X распределена по закону Бернулли с параметром p.

В итоге имеем следующее описание закона распределения:

X:

x_i	0	1
p_i	$1-p$	p

Здесь X – количество успехов в одном испытании; p – вероятность успеха; $q = 1 - p$ – вероятность неудачи.

Найдем числовые характеристики данной случайной величины.

$$MX = 0 \cdot (1-p) + 1 \cdot p = p;$$

$$MX^2 = 0^2 \cdot (1-p) + 1^2 \cdot p = p, \text{ откуда } DX = p - p^2 = p(1-p) = pq.$$

Таким образом, математическое ожидание и дисперсия случайной величины, распределенной по закону Бернулли с вероятностью успеха p, будут следующими:

$$MX = p;\ DX = pq.$$

II. Биномиальное распределение с параметрами *n* и *p* – $B_{n,p}$

Пусть проводится n испытаний по схеме Бернулли с вероятностью успеха, равной p в каждом испытании. Если X – количество успехов в описанной серии испытаний, то говорят, что случайная величина X имеет биномиальное распределение с параметрами n и p. Случайная величина X принимает значения 0, 1, 2, …, m, …, n с вероятностями

$$p_m = P(X = m) = C_n^m p^m q^{n-m} \text{ (формула Бернулли).}$$

Легко заметить, что основное свойство ряда распределения здесь выполнено в соответствии с формулой бинома Ньютона:

$$\sum_{m=0}^{n} p_m = q^n + npq^{n-1} + C_n^2 p^2 q^{n-2} + ... + C_n^m p^m q^{n-m} + ... + p^n = (q+p)^n = 1^n = 1.$$

Таким образом, случайная величина задана корректно.

Математическое ожидание и дисперсия случайной величины X, распределенной по биномиальному закону будут следующими:

$$MX = np;\ DX = npq.$$

Модой случайной величины, распределенной по биномиальному закону, является целое число, удовлетворяющее неравенству:

$$np - q \le MoX \le np + p.$$

В итоге имеем следующее описание биномиального распределения:

X:	x_i	0	1	2	…	m	...	n
	p_i	q^n	npq^{n-1}	$C_n^2 p^2 q^{n-2}$		$C_n^m p^m q^{n-m}$		p^n

где X – количество успехов в серии испытаний; где n – число испытаний; p – вероятность успеха; $q = 1 - p$ – вероятность неудачи.

Числовые характеристики случайной величины:

$$MX = np,\ DX = npq,\ np - q \le MoX \le np + p.$$

III. Распределение Пуассона с параметром λ – Π_λ

Дискретная случайная величина X распределена в соответствии с законом Пуассона с параметром λ, если она принимает целые неотрицательные значения 0, 1, 2, …, m, … с вероятностями

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

Функция $P(m,\lambda) = \frac{\lambda^m}{m!} e^{-\lambda}$ называется ***функцией Пуассона***. Ее значения могут быть вычислены с помощью инженерного калькулятора или математических программ. Для некоторых m и λ значения функции $P(m,\lambda)$ приведены в табл. 1 в приложении к сборнику задач [4].

Ряд распределения для закона Пуассона выглядит так:

X:	x_i	0	1	2	…	m	…
	p_i	$e^{-\lambda}$	$\lambda e^{-\lambda}$	$\frac{\lambda^2}{2!}e^{-\lambda}$	…	$\frac{\lambda^m}{m!}e^{-\lambda}$	…

Основное свойство ряда распределения выполнено, что легко подтвердить, используя разложение в ряд Маклорена для экспоненты:

$$\sum_{m=0}^{n} p_m = e^{-\lambda} + \lambda e^{-\lambda} + \frac{\lambda^2}{2!}e^{-\lambda} + \ldots + \frac{\lambda^m}{m!}e^{-\lambda} + \ldots =$$

$$= e^{-\lambda}\left(1 + \lambda + \frac{\lambda^2}{2!} + \ldots + \frac{\lambda^m}{m!} + \ldots\right) = e^{-\lambda} \cdot e^{\lambda} = 1.$$

Распределение Пуассона может рассматриваться как предельный случай биномиального распределения при бесконечно большом числе испытаний и ничтожно малой вероятности успеха. Закон Пуассона называют также «законом редких событий». С его помощью описывается число вызовов в системе массового обслуживания; число сбоев в работе механизма; количество опечаток в очередном номере журнала; число метеоритов, упавших на Землю в течение года и т. д.

Числовые характеристики случайной величины, распределенной по закону Пуассона.

$$MX = \lambda,\ DX = \lambda.$$

Среднее случайной величины равно значению параметра λ. Таким образом, λ – это среднее число происходящих в единицу времени событий.

IV. Геометрическое распределение с параметром *p* – G_p

Пусть испытания по схеме Бернулли с параметром p проводится до тех пор, пока одно из них не завершится успехом. Если X – номер первого успешного испытания, то говорят, что случайная величина X имеет геометрическое распределение с параметром p.

Случайная величина, распределенная по геометрическому закону, принимает значения 1, 2, …, m, … с вероятностями

$$p_m = P(X = m) = pq^{m-1}.$$

Ряд распределения для геометрического распределения имеет вид:

X:	x_i	1	2	3	…	m	…
	p_i	p	pq	pq^2	…	pq^{m-1}	…

Вероятности образуют геометрическую прогрессию с первым членом равным p и знаменателем $q = 1 - p$. Суммируя ряд, составленный из членов бесконечно убывающей геометрической прогрессии, проверим выполнение основного свойства ряда:

$$\sum_{m=0}^{n} p_m = p + pq + pq^2 + ... + pq^{m-1} + ... = \frac{p}{1-q} = \frac{p}{p} = 1.$$

Очевидно, что для случайной величины X, распределенной по геометрическому закону $MoX = 1$. Остальные числовые характеристики приведем без доказательства:

$$MX = \frac{1}{p},\ DX = \frac{q}{p^2}.$$

V. Гипергеометрическое распределение – $HG_{N,M,n}$

Пусть из совокупности N объектов, среди которых M обладают некоторым свойством, наудачу извлекается n объектов. Если X – количество объектов, обладающих свойством, среди извлеченных, то говорят, что X имеет гипергеометрическое распределение с параметрами N, M, n.

Случайная величина X принимает значения 1, 2, …, $\min(M, n)$. В соответствии с классическим определением вероятности и формулами комбинаторики

$$p_m = P(X = m) = \frac{C_M^m C_{N-M}^{n-m}}{C_N^n}.$$

Основное свойство ряда распределения и числовые характеристики для гипергеометрического распределения примем без доказательства.

$$MX = \frac{nM}{N},\ DX = \frac{nM}{N-1}\left(1 - \frac{M}{N}\right)\left(1 - \frac{n}{N}\right).$$

VI. Равномерное распределение на отрезке [*a*; *b*] – $U_{a,b}$

Случайная величина X распределена по равномерному на отрезке $[a,b]$ закону, если она имеет равные шансы попасть в любую точку этого отрезка.

Плотность распределения данной случайной величины постоянна на отрезке и равна нулю вне отрезка:

$$f(x)=\begin{cases}\dfrac{1}{b-a}, & \text{при } a\le x\le b;\\ 0, & \text{при } x<a \text{ или } x>b.\end{cases}$$

Непрерывным аналогом основного свойства ряда распределения являются свойства 1 и 2 плотности распределения (п. 2.4). Очевидно, что приведенная выше функция $f(x)$ принимает неотрицательные значения и интеграл от нее по всей числовой оси равен 1 (рис. 8).

Найдем функцию распределения.

При $x<a$ имеем $F(x)=\int\limits_{-\infty}^{x} f(t)dt=\int\limits_{-\infty}^{x} 0dt=0$.

При $a\le x\le b$:

$$F(x)=\int\limits_{-\infty}^{x} f(t)dt=\int\limits_{-\infty}^{a} 0dt+\int\limits_{a}^{x}\frac{1}{b-a}dt=0+\frac{1}{b-a}\cdot t\Big|_{a}^{x}=\frac{x-a}{b-a}.$$

При $x>b$: $F(x)=\int\limits_{-\infty}^{x} f(t)dt=\int\limits_{a}^{b}\frac{1}{b-a}dt=0+\frac{1}{b-a}\cdot t\Big|_{a}^{b}+0=\frac{b-a}{b-a}=1$.

В итоге функция распределения выглядит следующим образом:

$$F(x)=\begin{cases}0, & \text{при } x<a;\\ \dfrac{x-a}{b-a}, & \text{при } a\le x\le b;\\ 1, & \text{при } x>b.\end{cases}$$

Графики плотности распределения и функции распределения для равномерного закона даны на рис. 8 и 9.

Числовые характеристики случайной величины, распределенной по равномерному закону, таковы:

$$MX = \frac{b+a}{2},\ DX = \frac{(b-a)^2}{12}.$$

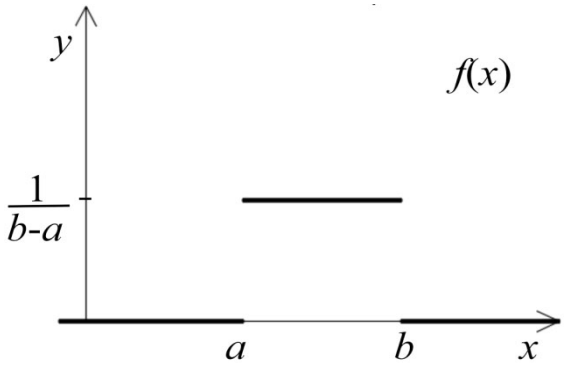


Рис. 8. Плотность равномерного распределения

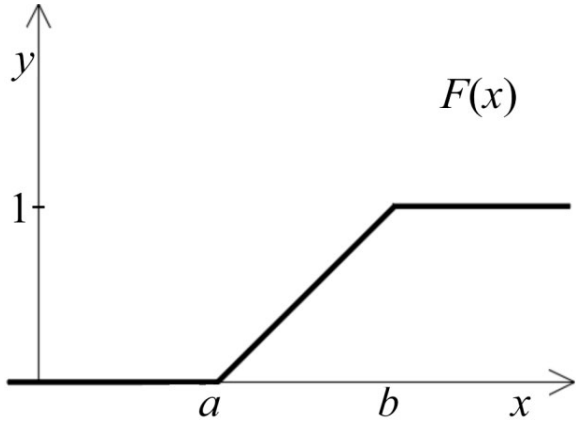


Рис. 9. Функция равномерного распределения

VII. Показательное распределение с параметром λ – E_λ

Случайная величина X распределена по показательному закону с параметром $\lambda > 0$, если ее плотность распределения и функция распределения имеют вид:

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{при } x \geq 0; \\ 0, & \text{при } x < 0. \end{cases} \qquad F(x) = \begin{cases} 1 - e^{-\lambda x}, & \text{при } x \geq 0; \\ 0, & \text{при } x < 0. \end{cases}$$

Графики плотности распределения и функции распределения для показательного распределения изображены на рис. 10 и 11.

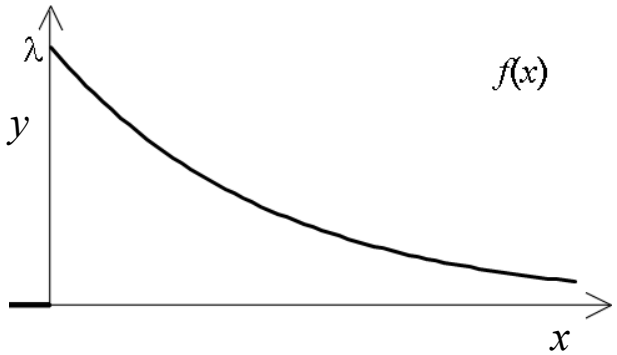


Рис. 10. Плотность показательного распределения

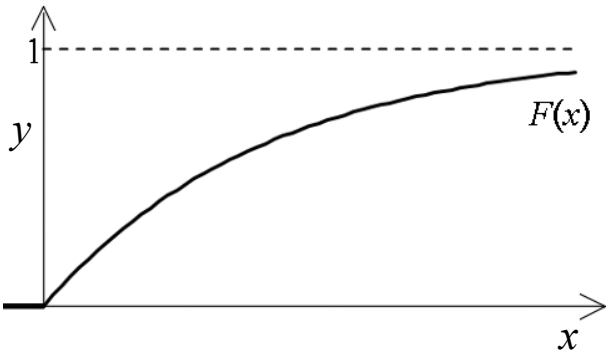


Рис. 11. Функция показательного распределения

Числовые характеристики случайной величины, распределенной по закону Пуассона, будут следующими:

$$MX = \frac{1}{\lambda};\ DX = \frac{1}{\lambda^2};\ \sigma_X = \frac{1}{\lambda}.$$

Показательное распределение связано с распределением Пуассона и описывает время ожидания редкого события (время между редкими событиями).

VIII. Нормальное распределение с параметрами *a* и σ^2 – N_{a,σ^2}

Случайная величина X имеет нормальный закон распределения (***закон Гаусса***) с параметрами a и σ^2, если ее плотность распределения задана функцией

$$\varphi_{a,\sigma^2}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-a)^2}{2\sigma^2}}, \text{ где } x \in (-\infty; +\infty).$$

График этой функции представляет собой кривую в форме колокола (кривая Гаусса – рис. 12), которая достигает максимума в точке $x = a$. Это значение является математическим ожиданием (а также модой и медианой) данной случайной величины.

Случайная величина, распределенная по нормальному закону, чаще принимает значения, близкие к среднему, и реже – отличные от среднего. Нормальным законом описываются многие природные и социальные явления – рост и вес человека, стоимость товара на рынке и т. д. [3].

Функция распределения нормальной случайной величины дается интегралом:

$$\Phi_{a,\sigma^2}(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(t-a)^2}{2\sigma^2}} dt.$$

Данный интеграл является неберущимся. Его значения могут быть получены с помощью математических программ или из специальных таблиц.

Числовые характеристики нормально распределенной случайной величины совпадают с ее параметрами:

$$MX = a;\ DX = \sigma^2;\ \sigma_X = \sigma;\ MoX = a;\ MeX = a.$$

Стандартное нормальное распределение имеет нулевое математическое ожидание и единичную дисперсию, $a = 0$; $\sigma^2 = 1$, откуда $\sigma = 1$. Его плотность и функция распределения имеют вид:

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}; \ \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt.$$

Приближенные значения функции $\varphi(x)$ даны в табл. 2 в приложении к сборнику задач [4] для $x \in [0;5]$ с шагом в 0,01. Целые и десятые доли x даны в строках, сотые в столбцах. Например, для того чтобы определить значение функции в точке 2,45, нужно найти строку с отметкой 2,4 и столбец с отметкой 5. На их пересечении находится число $\varphi(2,45) \approx 0,01984$. Функция $\varphi(x)$ является четной, следовательно $\varphi(-x) = \varphi(x)$. Для $x > 5$ значения $\varphi(x)$ очень близки к нулю.

Для вычисления значений функции распределения используется ***интеграл Лапласа*** $\widetilde{\Phi}(x) = \frac{1}{\sqrt{2\pi}} \int_{0}^{x} e^{-\frac{t^2}{2}} dt$, значения которого даны в табл. 3 [4]. В отличие от функции распределения, интеграл Лапласа имеет интервал интегрирования $[0;x]$, а не $(-\infty;x]$.

В силу того, что площадь под графиком плотности распределения $\varphi(x)$ равна 1, а половина этой площади (интеграл по интервалу $(-\infty;0]$) равна 0,5, значения функции распределения связаны с интегралом Лапласа следующим образом (рис. 13):

$$\Phi(x) = 0,5 + \widetilde{\Phi}(x).$$

Интеграл Лапласа является нечетной функцией, $\widetilde{\Phi}(-x) = -\widetilde{\Phi}(x)$. Для $x > 5$ его значения считаются приближенно равными 0,5.

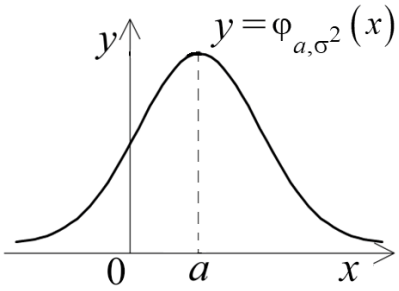


Рис. 12. Плотность нормального распределения

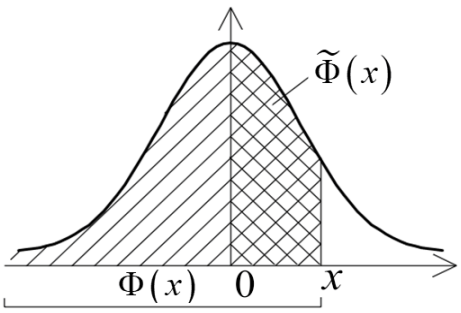


Рис. 13. Стандартное нормальное распределение

Теорема 8. Если случайная величина X распределена в соответствии с законом N_{a,σ^2}, то случайная величина $\dfrac{X-a}{\sigma}$ имеет распределение $N_{0,1}$.

Таким образом, для того чтобы найти значения плотности распределения и функции распределения нормально распределенной случайной величины с заданными параметрами a и σ^2, необходимо использовать следующие формулы:

$$\varphi_{a,\sigma^2}(x)=\varphi\left(\frac{x-a}{\sigma}\right); \qquad \widetilde{\Phi}_{a,\sigma^2}(x)=\widetilde{\Phi}\left(\frac{x-a}{\sigma}\right). \tag{1}$$

Формулы (2) и (3) позволяют найти вероятность попадания случайной величины в заданный интервал и вероятность того, что случайная величина отклонится от своего среднего значения не более чем на заданную величину d (вероятность попадания в интервал $[a-d;a+d]$).

$$P(x_1 \le X \le x_2)=\widetilde{\Phi}\left(\frac{x_2-a}{\sigma}\right)-\widetilde{\Phi}\left(\frac{x_1-a}{\sigma}\right); \tag{2}$$

$$P(|X-a|\le d)=2\widetilde{\Phi}\left(\frac{d}{\sigma}\right). \tag{3}$$

Заметим, что в формулах (2) и (3) неравенство может быть строгим – это не повлияет на вероятность в силу теоремы 7, так как случайная величина, распределенная по нормальному закону, является непрерывной.

С вероятностью, близкой к 1, значения такой случайной величины отличаются от среднего не более чем на три стандартных отклонения (***правило трех сигм***). Иными словами, значения случайной величины,

распределенной по нормальному закону с заданными параметрами a и σ^2, практически достоверно лежит в интервале

$$\left(a-3\sigma,\, a+3\sigma\right). \qquad (4)$$

Выясним смысл слов «практически достоверно». В соответствии с формулой (3)

$$P\left(\left|X-a\right|\le 3\sigma\right)=2\widetilde{\Phi}\left(\frac{3\sigma}{\sigma}\right)=2\widetilde{\Phi}\left(3\right)\approx\left|\text{см. табл. 3}\right|\approx 2\cdot 0,49865\approx 0,9973.$$

Таким образом, вероятность того, что значение нормально распределенной случайной величины выйдет за пределы трех стандартных отклонений, очень мала, и составляет 0,0027.

Пример 16. Случайная величина X распределена по нормальному закону с $a=2$ и $\sigma^2=9$. Найти вероятности следующих событий:

а) X примет значение меньше 1; **б)** X примет значение больше 3,5;

в) X примет значение из интервала $\left[0;5\right]$;

г) X отклонится от своего среднего значения не более чем на 0,5.

д) Найти интервал, в котором практически достоверно лежат значения случайной величины X.

По условию задачи $a=2$ и $\sigma=3$.

а) Требуется найти $P\left(X<1\right)$. По определению это значение функции распределения в точке $x=1$. По формуле (1):

$$P\left(X<1\right)=\Phi_{2,9}\left(1\right)=\frac{1}{2}+\widetilde{\Phi}\left(\frac{1-2}{3}\right)=\frac{1}{2}+\widetilde{\Phi}\left(-\frac{1}{3}\right)=\left|\begin{matrix}\text{так как }\widetilde{\Phi}\\ \text{нечетная}\end{matrix}\right|=\frac{1}{2}-\widetilde{\Phi}\left(\frac{1}{3}\right)=$$

$$\approx\left|\begin{matrix}\text{округляем } x \text{ до}\\ \text{второго знака}\end{matrix}\right|\approx\frac{1}{2}-\widetilde{\Phi}\left(0,33\right)\approx\left|\text{из табл. 3}\right|\approx 0,5-0,12930=0,3807.$$

б) Перейдем к противоположному событию:

$$P\left(X>3,5\right)=1-P\left(X\le 3,5\right)=1-\Phi_{2,9}\left(3,5\right)=1-\left(\frac{1}{2}+\widetilde{\Phi}\left(\frac{3,5-2}{3}\right)\right)=$$

$$=\frac{1}{2}-\widetilde{\Phi}\left(\frac{1,5}{3}\right)=\frac{1}{2}-\widetilde{\Phi}\left(0,5\right)\approx 0,5-0,19146=0,30854.$$

в) Воспользуемся формулой (2):

$$P\left(0 \le X \le 5\right) = \widetilde{\Phi}\left(\frac{5-2}{3}\right) - \widetilde{\Phi}\left(\frac{0-2}{3}\right) = \widetilde{\Phi}\left(1\right) - \widetilde{\Phi}\left(-\frac{2}{3}\right) \approx \widetilde{\Phi}\left(1\right) + \widetilde{\Phi}\left(0,67\right) \approx$$

$$\approx 0,34134 + 0,24857 = 0,58991.$$

г) Воспользуемся формулой (3):

$$P\left(\left|X-2\right| \le 0,5\right) = 2\widetilde{\Phi}\left(\frac{0,5}{3}\right) = 2\widetilde{\Phi}\left(\frac{1}{6}\right) \approx 2\widetilde{\Phi}\left(0,17\right) \approx 2 \cdot 0,06749 = 0,13498.$$

д) В соответствии с формулой (4) X практически достоверно лежит в интервале

$$\left(2 - 3 \cdot 3; 2 + 3 \cdot 3\right) \text{ или } \left(-7;11\right).$$

3. ПОВТОРНЫЕ НЕЗАВИСИМЫЕ ИСПЫТАНИЯ

3.1. Схема Бернулли

Пусть проводится серия из n испытаний, вероятность наступления некоторого события (вероятность успеха) в каждом из которых равна p. Как известно из п. 2.6., число успешных испытаний распределено по биномиальному закону, и вероятность того, что успешными будут ровно m испытаний, дается формулой Бернулли:

$$P_n(m) = C_n^m p^m q^{n-m}, \text{ где } m = 0,1,...,n \text{ и } q = 1 - p.$$

На практике часто требуется вычислять такие вероятности при больших значениях m и n. Например, пусть требуется найти вероятность 200 успехов в 300 испытаниях. По формуле Бернулли:

$$P_{300}(200) = C_{300}^{200} p^{200} q^{100} = \frac{300!}{200!100!} p^{200} q^{100}.$$

Вычислить требуемое значение по формуле Бернулли технически невозможно. В таких случаях принято пользоваться формулами, которые позволяют найти значения $P_n(m)$ приближенно.

В частности, при очень малых значениях p биномиальное распределение очень близко к распределению Пуассона.

3.2. Формула Пуассона

Теорема 9 (теорема Пуассона). Пусть число испытаний n по схеме Бернулли велико, а вероятность успеха в одном испытании p очень мала, и мало также значение $\lambda = np$. Тогда $P_n(m)$ приближенно вычисляется по ***формуле Пуассона***:

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

Значения функции $P(m,\lambda)$ соответствуют вероятностям распределения Пуассона (см. п. 2.6) и могут быть взяты из табл. 1 [4].

Формула Пуассона может быть применена при $n \geq 1000$ и $\lambda \leq 10$, однако возможно ее использование и при меньшем числе испытаний. Например, если $n \geq 100$, то $\lambda \leq 5$; если $n \geq 20$, то $\lambda \leq 3$ [1].

Пример 17. В фирме работает 1460 сотрудников. Какова вероятность того, что 1 июня является днем рождения ровно пяти сотрудников? Не менее чем троих сотрудников?

Число испытаний $n = 1460$. Вероятность того, что 1 июня (как и любой другой день) является днем рождения наудачу выбранного человека, равна $p = \frac{1}{365}$. Следовательно, $\lambda = np = \frac{1460}{365} = 4$.

Так как $n \geq 1000$ и $\lambda \leq 10$, применять формулу Пуассона можно.

$$P_{1460}(5) \approx P(5,4) \approx |\text{из табл. 1}| \approx 0{,}15629.$$

Поскольку «не менее трех» означает «от 3 до 1460», то для ответа на второй вопрос задачи требуется рассмотреть противоположное событие и применить теорему 3:

$$P(X \geq 3) = 1 - P(X < 3) = 1 - \left(P(0) + P(1) + P(2)\right) \approx$$
$$\approx 1 - P(0,4) - P(1,4) - P(2,4) \approx 1 - 0{,}01832 - 0{,}07326 - 0{,}14653 = 0{,}23811.$$

3.3. Формулы Муавра – Лапласа

Если в серии испытаний по схеме Бернулли велико не только число испытаний *n*, но и величина *np*, то формула Пуассона использована быть не может в силу слишком большой погрешности. В этом случае для вычисления вероятностей $P_n(m)$ следует применять формулы Муавра – Лапласа – *локальную* или *интегральную*, согласно которым биномиальное распределение в такой ситуации близко к нормальному.

При этом локальная теорема применяется, если нужно найти вероятность ровно *m* успехов в *n* испытаниях, а интегральная – если вычисляется вероятность получить число успехов, заключенное в пределах от m_1 до m_2.

Теорема 10 (локальная теорема Муавра – Лапласа). Если число испытаний n по схеме Бернулли велико (а именно, если $npq \geq 20$), то $P_n(m)$ приближенно вычисляется по ***локальной формуле Муавра – Лапласа***:

$$P_n(m) \approx \frac{\varphi(x)}{\sqrt{npq}}, \text{ где } x = \frac{m - np}{\sqrt{npq}}.$$

Здесь $\varphi(x)$ – плотность стандартного нормального распределения. Ее значения могут быть взяты в табл. 2 [4]. При этом следует помнить, что $\varphi(-x) = \varphi(x)$.

Теорема 11 (интегральная теорема Муавра – Лапласа). Если число испытаний n по схеме Бернулли велико (а именно, если $npq \geq 20$), то вероятность того, что число успехов в n испытаниях заключено в пределах от m_1 до m_2 приближенно вычисляется по ***интегральной формуле Муавра – Лапласа***:

$$P(m_1 \leq X \leq m_2) \approx \widetilde{\Phi}(x_2) - \widetilde{\Phi}(x_1), \text{ где } x_1 = \frac{m_1 - np}{\sqrt{npq}},\ x_2 = \frac{m_2 - np}{\sqrt{npq}}.$$

Здесь $\widetilde{\Phi}(x)$ – интеграл Лапласа, значения которого находятся по табл. 3 [4]. При этом следует помнить, что $\widetilde{\Phi}(-x) = -\widetilde{\Phi}(x)$.

Пример 18. Известно, что в среднем 8 из 10 человек имеют сотовый телефон. Найти вероятность того, что из 400 наудачу выбранных людей сотовые телефоны будут иметь: **а)** 300 человек; **б)** от 310 до 360 человек.

Из условия задачи $n = 400$, $p = 0{,}8$. Отсюда $q = 1 - 0{,}8 = 0{,}2$ и $npq = 400 \cdot 0{,}8 \cdot 0{,}2 = 64 \geq 20$. Следовательно, можно применить формулы Муавра – Лапласа.

а) Используем локальную формулу Муавра – Лапласа.

$$x = \frac{300 - 400 \cdot 0{,}8}{\sqrt{64}} = \frac{300 - 320}{8} = -2{,}5; \qquad \varphi(-2{,}5) = \varphi(2{,}5) \approx 0{,}01753;$$

Согласно теореме 10, $P_{400}(300) \approx \dfrac{0{,}01753}{8} \approx 0{,}0022$.

б) Используем интегральную формулу Муавра – Лапласа.

$$x_1 = \frac{310-320}{8} = \frac{-10}{8} = -1,25\,;\ x_2 = \frac{360-320}{8} = \frac{40}{8} = 5\,.$$

Согласно теореме 11,

$$P(310 \le X \le 360) \approx \widetilde{\Phi}(5) - \widetilde{\Phi}(-1,25) \approx \widetilde{\Phi}(5) + \widetilde{\Phi}(1,25) \approx$$

$$\approx 0,5 + 0,39435 = 0,89435\,.$$

$\widetilde{\Phi}(5)$ в данном случае можно считать равным 0,5, так как приведенное в таблице значение отличается от 0,5 только в седьмом знаке. При решении большинства задач достаточно точности до пятого знака.

4. СИСТЕМЫ ДВУХ СЛУЧАЙНЫХ ВЕЛИЧИН

4.1. Дискретные двумерные случайные величины

Возможна ситуация, когда результат испытания является парой чисел (X,Y), представляющих собой значения двух различных случайных величин. Случайные величины X и Y, рассмотренные в совокупности, называются ***системой двух случайных величин,*** или ***двумерной случайной величиной***.

В общем случае система нескольких случайных величин задается случайным вектором $\overline{X}=(X_1,X_2,...X_n)$ и называется ***n-мерной*** или ***многомерной случайной величиной***.

Пример 19. Примером двумерной случайной величины могут служить рост и вес выбранного наудачу человека.

В качестве примера многомерной случайной величины рассмотрим успеваемость студента в течение сессии. Такая случайная величина представляет собой вектор оценок по различным дисциплинам, некоторые из которых производятся по четырехбалльной системе, некоторые могут принимать значения «зачтено» и «не зачтено».

Для дискретной двумерной случайной величины закон распределения задается таблицей, где строки соответствуют значениям, которые принимает X, столбцы – значениям, которые принимает Y. В ячейках таблицы располагаются $p_{ij}=P(x_i,y_j)=P(\{X=x_i\}\cap\{Y=y_j\})$ – вероятности реализации каждой пары значений (X,Y).

$X \backslash Y$	y_1	y_2	$\dots$	y_j	$\dots$	y_m
x_1	p_{11}	p_{12}	$\dots$	p_{1j}	$\dots$	p_{1m}
x_2	p_{21}	p_{22}	$\dots$	p_{2j}	$\dots$	p_{2m}
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
x_i	p_{i1}	p_{i2}	$\dots$	p_{ij}	$\dots$	p_{im}
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
x_n	p_{n1}	p_{n2}	$\dots$	p_{nj}	$\dots$	p_{nm}

Основное свойство ряда распределения здесь имеет вид:

$$\sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij} = 1.$$

По таблице двумерной случайной величины можно получить ряды распределения отдельно для X и Y. Случайная величина X принимает значения x_i с вероятностями $p_i = \sum_{j=1}^{m} p_{ij}$ при $i = 1, 2, ..., n$, а случайная величина Y значения y_j с вероятностями $q_j = \sum_{i=1}^{n} p_{ij}$ при $j = 1, 2, ..., m$.

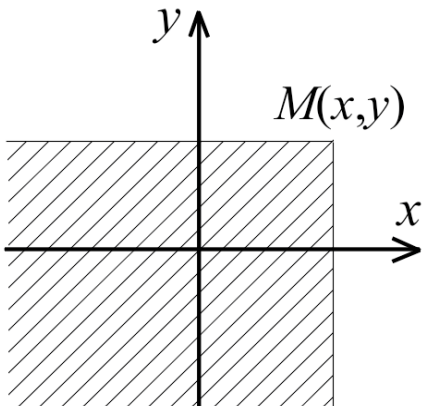


Рис. 14. Совместная функция распределения

Функция распределения двумерной случайной величины является функцией двух переменных $F(x, y) = P(\{X < x\} \cap \{Y < y\})$ и равна вероятности попадания случайной точки (X, Y) в область на рис. 14. $F(x, y)$ называется также ***совместной функцией распределения*** случайных величин X и Y.

Для дискретной двумерной случайной величины функция распределения постоянна в областях, ограниченных прямыми $x = x_i$ и $y = y_j$. Ее значения в каждой из областей находятся по формуле

$$F(x, y) = \sum_{x_i < x} \sum_{y_j < y} p_{ij},$$

а график представляет собой ступенчатую поверхность.

Две случайные величины X и Y являются независимыми тогда и только тогда, когда их совместная функция распределения равна произведению функций распределения для X и Y в отдельности. То же верно и для вероятностей каждой пары значений в силу теоремы умножения вероятностей для независимых событий (теорема 5, п. 1.6).

Сформулируем критерий независимости дискретных случайных величин.

Теорема 12. X и Y независимы $\Leftrightarrow$ $F(x,y)=F_1(x)\cdot F_2(y)$ $\Leftrightarrow$ $p_{ij}=p_i\cdot q_j$.

4.2. Условные законы распределения

Пусть задана двумерная случайная величина (X,Y). Зафиксируем значение $Y=y_j$. Случайная величина X при этом будет подчинена ***условному закону распределения для*** X ***при*** $Y=y_j$. Такая случайная величина обозначается $X_{Y=y_j}$. Ее значения по-прежнему равны x_i, где $i=1,2,...,n$, а их вероятности находятся по формуле

$$p_i=\frac{p_{ij}}{P\left(Y=y_j\right)}=\frac{p_{ij}}{q_j}.$$

Аналогично задается ***условное распределение случайной величины*** Y ***при условии*** $X=x_i$, которая обозначается $Y_{X=x_i}$ и принимает значения y_j, $j=1,2,...,m$ с вероятностями

$$q_j=\frac{p_{ij}}{P\left(X=x_i\right)}=\frac{p_{ij}}{p_i}.$$

Математическое ожидание и дисперсия условных распределений называются ***условным математическим ожиданием*** и ***условной дисперсией*** и обозначаются M_YX, D_YX для случайной величины $X_{Y=y_j}$ и M_XY, D_XY для случайной величины $Y_{X=x_i}$.

4.3. Ковариация и коэффициент корреляции

Кроме средних и дисперсий случайных величин X и Y, двумерная случайная величина (X,Y) имеет совместные числовые характеристики, выражающие степень взаимной зависимости ее составляющих. Таковыми являются ковариация и коэффициент корреляции.

Ковариацией случайных величин X и Y называется математическое ожидание произведения отклонений этих случайных величин от своих средних значений:

$$K_{XY} = M\left[\left(X - MX\right)\left(Y - MY\right)\right].$$

Для дискретной двумерной случайной величины:

$$K_{XY} = \sum_{i=1}^{n}\sum_{j=1}^{m}\left(x_i - MX\right)\left(y_j - MY\right);$$

<u>Свойства ковариации:</u>

1) $K_{XX} = M\left[\left(X - MX\right)\left(X - MX\right)\right] = M\left(X - MX\right)^2 = DX$;

2) $K_{XY} = M\left(XY\right) - MX \cdot MY$;

3) Если X и Y независимы, то $M\left(XY\right) = MX \cdot MY$, следовательно $K_{XY} = 0$;

4) $\left|K_{XY}\right| \le \sigma_X \cdot \sigma_Y$, причем равенство в данной формуле достигается, когда случайные величины X и Y связаны линейно, то есть существуют постоянные a,b такие, что $Y = aX + b$.

Ковариация может рассматриваться как мера зависимости случайных величин, которая обращается в ноль, если эти случайные величины независимы и принимает максимально возможное значение $\sigma_X \cdot \sigma_Y = \sqrt{DX \cdot DY}$, если между ними существует детерминированная линейная связь.

Разделив ковариацию на величину $\sigma_X \cdot \sigma_Y$, получим более удобную числовую характеристику зависимости случайных величин, называемую ***коэффициентом корреляции***:

$$r_{XY} = \frac{K_{XY}}{\sigma_X \cdot \sigma_Y} = \frac{M\left(XY\right) - MX \cdot MY}{\sqrt{DX \cdot DY}}.$$

<u>Свойства коэффициента корреляции:</u>

1) $-1 \le r_{XY} \le 1$;

2) Если X и Y независимы, то $r_{XY} = 0$;

3) $r_{XX} = \dfrac{K_{XX}}{\sigma_X \cdot \sigma_X} = \dfrac{DX}{DX} = 1$;

4) $r_{XY} = \pm 1$ тогда и только тогда, когда между X и Y существует функциональная линейная связь, то есть можно указать постоянные a, b такие, что $Y = aX + b$; при этом $r_{XY} = 1$, если $a > 0$, $r_{XY} = -1$, если $a < 0$.

Коэффициент корреляции является нормированной величиной, характеризующей степень взаимосвязи случайных величин. Если $r_{XY} = 0$, то случайные величины X и Y называются ***некоррелированными***. Как следует из свойства 2, если X и Y независимы, то они некоррелированны. Обратное неверно. Между случайными величинами может существовать сильная нелинейная зависимость, выявить которую коэффициент корреляции не поможет.

Если $r_{XY} > 0$, то говорят, что X и Y ***положительно коррелированы***. В этом случае с ростом значений одной случайной величины, значения второй имеют тенденцию увеличиваться. Если $r_{XY} < 0$, то X и Y ***отрицательно коррелированы***, то есть при возрастании одной случайной величины, другая в среднем будет принимать меньшие значения.

Пример 20. Двумерная случайная величина (X, Y) задана таблицей.

$X \backslash Y$	0	1	2
2	0,2	0,15	0,1
4	0,1	0,2	0,25

Найти:

а) законы распределения для X и Y в отдельности;

б) сумму случайных величин $U = X + Y$;

в) вероятность события $2Y \geq X$;

г) условные законы распределения для X при $Y = 1$ и для Y при $X = 4$; условные математические ожидания и дисперсии;

д) ковариацию и коэффициент корреляции.

а) Случайная величина X принимает значения 2 и 4. Для того чтобы вычислить вероятность значения 2, необходимо просуммировать вероятности по первой строке таблицы; для 4 – по второй строке:

$$P(X = 2) = 0,2 + 0,15 + 0,1 = 0,45; \; P(X = 4) = 0,1 + 0,2 + 0,25 = 0,55.$$

Для случайной величины Y вероятности суммируются по столбцам:

$$P(Y=0)=0,2+0,1=0,3;\ P(Y=1)=0,15+0,2=0,35;$$
$$P(Y=2)=0,1+0,25=0,35.$$

В итоге законы распределения для X и Y задаются таблицами:

X :

x_i	2	4
p_i	0,45	0,55

Y :

y_j	0	1	2
q_j	0,3	0,35	0,35

б) для вычисления суммы случайных величин построим таблицу аналогично тому, как это было сделано при решении примера 18 в п. 2.1. с учетом того, что вероятности для каждой пары значений уже известны.

X	Y	$U=X+Y$	P
2	0	2	0,2
2	1	3	0,15
2	2	4	0,1
4	0	4	0,1
4	1	5	0,2
4	2	6	0,25

Случайная величина U принимает значения 2, 3, 4, 5, 6, причем значение 4 возможно в двух случаях – для пары (2, 2) и для пары (4, 0). Так как события, состоящие в появлении этих пар, являются несовместными, то по теореме 1:

$$P(U=4)=P\big((2,2)\cup(4,0)\big)=P(2,2)+P(4,0)=0,1+0,1=0,2.$$

Закон распределение для U выглядит следующим образом:

U :

$\boldsymbol{u_i}$	2	3	4	5	6
$\boldsymbol{p_i}$	0,2	0,15	0,2	0,2	0,25

в) По таблице распределения просуммируем вероятности всех пар (x_i, y_j), для которых выполнено условие $y_j \ge x_i$:

$$P(2Y \ge X)=P\big((2,1)\cup(2,2)\cup(4,2)\big)=|\text{теор.1}|=0,15+0,1+0,25=0,5.$$

г) Найдем закон распределения X при $Y=1$. Как было показано ранее, $P(Y=1)=0,35$. Вычислим вероятности всех значений X при этом условии:

$$P(2)=\frac{0,15}{0,35}=\frac{3}{7};\ P(4)=\frac{0,2}{0,35}=\frac{4}{7}.$$

Аналогично, $P(X=4)=0,55$, а вероятности значений Y:

$$P(0)=\frac{0,1}{0,55}=\frac{2}{11};\ P(2)=\frac{0,2}{0,55}=\frac{4}{11};\ P(2)=\frac{0,25}{0,55}=\frac{5}{11}.$$

В итоге условные законы распределения задаются таблицами:

$X_{Y=1}$:

x_i	2	4
p_i	$\frac{3}{7}$	$\frac{4}{7}$

$Y_{X=4}$:

y_j	0	1	2
q_j	$\frac{2}{11}$	$\frac{4}{11}$	$\frac{5}{11}$

д) Для вычисления ковариации используем формулу из свойства 2:

$$K_{XY}=M(XY)-MX\cdot MY.$$

Законы распределения для X и Y были найдены в п. а). Найдем их математические ожидания:

$$MX=2\cdot 0,45+4\cdot 0,55=3,1;\qquad MY=0\cdot 0,3+1\cdot 0,35+2\cdot 0,35=1,05.$$

Закон распределения случайной величины XY строить не обязательно. Значение $M(XY)$ может быть найдено как $M(XY)=\sum_{i=1}^{n}\sum_{j=1}^{m}x_i y_j p_{ij}$. То есть,

$$M(XY)=2\cdot 0\cdot 0,2+2\cdot 1\cdot 0,15+2\cdot 2\cdot 0,1+4\cdot 0\cdot 0,1+4\cdot 1\cdot 0,2+4\cdot 2\cdot 0,25=3,5.$$

Вычислим ковариацию:

$$K_{XY}=3,5-3,1\cdot 1,05=0,245.$$

Для вычисления коэффициента корреляции необходимо найти дисперсии случайных величин X и Y:

$$MX^2=2^2\cdot 0,45+4^2\cdot 0,55=10,6;\ DX=10,6-3,1^2=0,99;$$

$$MY^2=0^2\cdot 0,3+1^2\cdot 0,35+2^2\cdot 0,35=1,75;\ DY=1,75-1,05^2=0,6475.$$

И, наконец, найдем коэффициент корреляции:

$$r_{XY} = \frac{K_{XY}}{\sqrt{DX \cdot DY}} = \frac{0,245}{\sqrt{1,75 \cdot 0,6475}} \approx 0,228.$$

5. ОСНОВНЫЕ ПОНЯТИЯ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

Математическая статистика, как и теория вероятностей, изучает случайные явления. Однако задачи, которые она решает, имеют иной характер. В теории вероятностей предполагается, что закон распределения исследуемых случайных величин известен или может быть найден в ходе решения задачи. В математической статистике закон распределения неизвестен или известен не полностью. Вместо этого имеются данные наблюдений – случайные величины, распределенные в соответствии с неизвестным нам законом распределения. Задача математической статистики – «восстановить» закон распределения по этим данным [1].

В математической статистике закон распределения никогда не может быть восстановлен точно. Можно только оценить тип распределения или его неизвестные параметры и проанализировать качество такой оценки.

5.1. Выборка и вариационный ряд

Пусть на практике наблюдается многократная реализация случайных величин, распределенных по одному и тому же закону, который необходимо определить. Множество всех возможных результатов наблюдений называется ***генеральной совокупностью***.

Для изучения свойств генеральной совокупности проводится некоторое количество испытаний, результат каждого из которых описывается числом X_i, где $i = 1, 2, ..., N$. Последовательность независимых одинаково распределенных случайных величин $X_1, X_2, ..., X_N$, полученная таким образом, называется ***выборкой*** объема N. Число X_i – ***i***-й ***элемент выборки***.

Выборка является единственным материалом, из которого можно извлечь информацию о генеральной совокупности. Чем больше объем выборки, тем точнее можно с ее помощью описать закон распределения. Важным также является вопрос о репрезентативности выборки – то есть о том, насколько хорошо она характеризует генеральную совокупность.

Пример 21. Даны результаты бросания кубика:

6; 5; 2; 6; 2; 3; 4; 1; 2; 2.

Данные представляют собой выборку объема 10 ($N = 10$).

Вариационным рядом называется выборка, упорядоченная по возрастанию. Элементы вариационного ряда обозначаются $X_{(1)}, X_{(2)}, ..., X_{(N)}$. При этом $X_{(1)} = \min\limits_{i=1,...,N} \{X_i\}$, $X_{(N)} = \max\limits_{i=1,...,N} \{X_i\}$.

Пример 21 (продолжение). Вариационный ряд имеет вид:

1; 2; 2; 2; 2; 3; 4; 5; 6; 6.

Здесь $X_{(1)} = 1$; $X_{(2)} = X_{(3)} = X_{(4)} = X_{(5)} = 2$; … ; $X_{(N)} = X_{(10)} = 6$.

Данные выборки часто представляются в сгруппированном виде. Пусть элементы выборки принимают k различных значений $x_1, x_2, ..., x_k$ (обычно k заметно меньше N). Значения x_i называются ***вариантами*** (ед. ч. – ***варианта***). Они являются константами, а не случайными величинами. Количество элементов выборки, значение которых равно x_i называется ***частотой***, соответствующей варианте x_i и обозначается n_i. Очевидно, что $n_1 + n_2 + ... + n_k = N$.

Пример 21 (продолжение). Сгруппируем выборку по вариантам:

x_i	1	2	3	4	5	6
n_i	1	4	1	1	1	2

Группировка по вариантам используется, если элементы выборки распределены в соответствии с некоторым дискретным законом. Если дана выборка из непрерывного распределения, то ее элементы, как правило, различны и могут быть сгруппированы по интервалам. Промежуток $[a;b]$, содержащий все значения, которые принимают элементы, разбивается на k интервалов точками $a = x_0 < x_1 < x_2 < ... < x_{k-1} < x_k = b$. Обычно эти интервалы имеют одинаковую длину, но это не всегда так. Группировка по интервалам может производиться уже на этапе сбора данных.

Частотой n_i здесь будет называться количество элементов выборки, попавших в i-й интервал $[x_{i-1}, x_i)$. Все интервалы являются полуоткрытыми, кроме последнего, $[x_{k-1}, x_k]$.

Долей (***относительной частотой***), соответствующей i-му интервалу, называется величина $w_i = \dfrac{n_i}{N}$ (доля элементов выборки, попавших в i-й интервал).

Накопленной частотой, $n_i^{\text{нак}}$ называется количество элементов выборки, значение которых меньше x_i. Фактически это сумма частот по интервалам с первого по i-й.

Накопленная доля – величина $w_i^{\text{нак}} = \dfrac{n_i^{\text{нак}}}{N}$.

Пример 22. Имеются данные об объеме производства по сравнению с предыдущим годом для 100 предприятий одной отрасли. Курсивом выделены исходные данные. Вычислим остальные величины.

i	$[x_{i-1}, x_i)$	n_i	w_i	$n_i^{\text{нак}}$	$w_i^{\text{нак}}$
1	***95 – 100 %***	***8***	0,08	**8**	0,08
2	***100 – 105 %***	***28***	0,28	8+28=**36**	0,36
3	***105 – 110 %***	***36***	0,36	36+36=**72**	0,72
4	***110 – 115 %***	***24***	0,24	72+24=**96**	0,96
5	***больше 115 %***	***4***	0,04	96+4=**100**	1,00
	$\sum$	100	1	–	–

Рекомендованное число интервалов для группировки выборки объема N может быть определено по ***формуле Стёрджеса***: $m = 1 + \lfloor \log_2 N \rfloor$, где $\lfloor x \rfloor$ – целая часть числа x. Так как $\log_2 N = \dfrac{\lg N}{\lg 2} = \log_2 10 \lg N \approx 3{,}322 \lg N$, формула Стёрджеса может иметь вид $m = 1 + \lfloor 3{,}322 \lg N \rfloor$.

Пример 22 (продолжение). Для рассматриваемого в примере случая $N = 100$ $m = 1 + \lfloor 3{,}322 \lg 100 \rfloor = 1 + \lfloor 3{,}322 \cdot 2 \rfloor = 7$.

5.2. Теоретическая и эмпирическая функция распределения

Пусть задана выборка $X_1, X_2, ..., X_N$ объема N. Закон распределения случайных величин X_i в задачах математической статистики неизвестен. Он называется ***теоретическим законом распределения***. Функция

распределения случайных величин X_i, $F(x)=P(X_i<x)$ называется ***теоретической функцией распределения***. Если существует плотность распределения $f(x)=F'(x)$, то она называется ***теоретической плотностью распределения***.

Полностью восстановить теоретический закон распределения нельзя, однако можно дать ему приближенную оценку.

Эмпирической функцией распределения называется функция $F^*(x)=w_i^{\text{нак}}$ при $x\in[x_{i-1},x_i)$. График функции $F^*(x)$ распределения называется ***кумулятивной кривой***. Функция $f^*(x)=w_i$ называется ***гистограммой*** или ***полигоном частот***. Фактически это эмпирическая плотность распределения или эмпирический ряд распределения.

Пример 22 (продолжение). Эмпирическая функция распределения и гистограмма по данным примера 22, изображены на рис. 15 и 16.

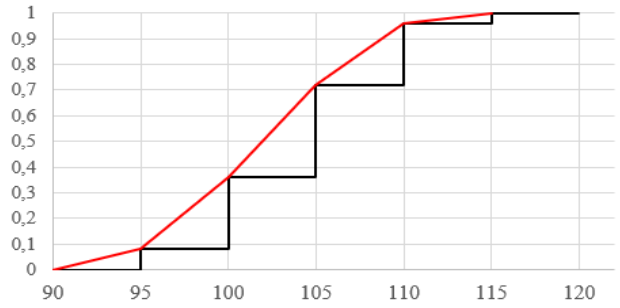


Рис. 15. Эмпирическая функция распределения для примера 22

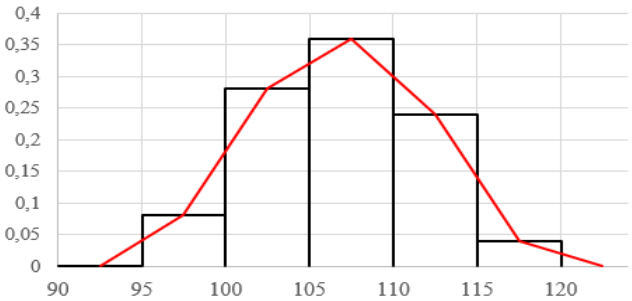


Рис. 16. Гистограмма для примера 22

Пример 21 (продолжение). Дополним таблицу и построим эмпирическую функцию распределения и гистограмму (рис. 17 и 18).

x_i	**1**	**2**	**3**	**4**	**5**	**6**
n_i	**1**	**4**	**1**	**1**	**1**	**2**
w_i	0,1	0,4	0,1	0,1	0,1	0,2
$n_i^{\text{нак}}$	1	5	6	7	8	10
$w_i^{\text{нак}}$	0,1	0,5	0,6	0,7	0,8	1,0

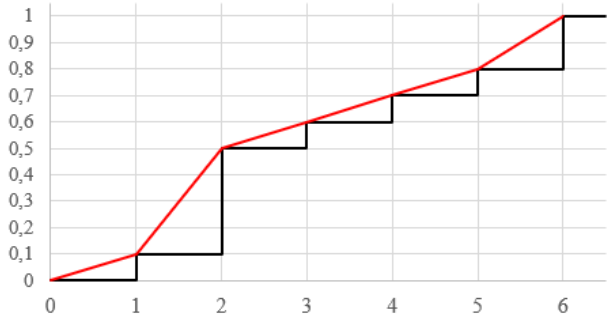


Рис. 17. Эмпирическая функция распределения для примера 21

Рис. 18. Гистограмма для примера 21

По внешнему виду гистограммы можно выдвинуть гипотезу о теоретическом законе распределения. Например, рис. 16 позволяет предположить, что элементы выборки распределены по нормальному закону, а в случае рис. 18 – по дискретному равномерному.

Проверке таких гипотез посвящен раздел 10 данного издания.

5.3. Числовые характеристики выборки

Числовой характеристикой выборки (***выборочной характеристикой***), или ***статистикой*** называется функция всех элементов выборки. Рассмотрим наиболее важные статистики.

1) ***Выборочное среднее***:

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i .$$

Для выборки, сгруппированной по вариантам:

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{k} x_i n_i \text{, где } k \text{ – количество вариант.}$$

Для выборки, сгруппированной по интервалам:

$\overline{X} = \frac{1}{N}\sum_{i=1}^{k} \overline{x}_i n_i$; k – количество интервалов; $\overline{x}_i$ – середина i-го интервала.

2) k ***-й выборочный момент***:

$$\overline{X^k} = \frac{1}{N}\sum_{i=1}^{N} X_i^{\,k} .$$

Для выборки, сгруппированной по вариантам:

$$\overline{X^k} = \frac{1}{N}\sum_{i=1}^{k} x_i^{\,k} n_i \text{, где } k \text{ – количество вариант.}$$

Для выборки, сгруппированной по интервалам:

$\overline{X} = \frac{1}{N}\sum_{i=1}^{k} \overline{x}_i^{\,k} n_i$; k – количество интервалов; $\overline{x}_i$ – середина i-го интервала.

Наиболее важное значение имеет второй выборочный момент:

$$\overline{X^2} = \frac{1}{N}\sum_{i=1}^{N} X_i^{\,2} \text{ или } \overline{X^2} = \frac{1}{N}\sum_{i=1}^{k} x_i^{\,2} n_i \text{ или } \overline{X} = \frac{1}{N}\sum_{i=1}^{k} \overline{x}_i^{\,2} n_i .$$

3) ***Выборочная дисперсия***:

$$\sigma^{2*} = \frac{1}{N}\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2 \text{ или } \sigma^{2*} = \overline{X^2} - \left(\overline{X}\right)^2 .$$

Исправленная (***несмещенная***) ***выборочная дисперсия***:

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2 \text{ или } s^2 = \frac{N}{N-1}\sigma^{2*} .$$

Выборочное среднее квадратическое отклонение: $s = \sqrt{s^2}$.

Пример 21 (продолжение). Найдем статистики выборки:

x_i	1	2	3	4	5	6
n_i	1	4	1	1	1	2

$$\overline{X} = \frac{1}{10}\left(1\cdot 1 + 2\cdot 4 + 3\cdot 1 + 4\cdot 1 + 5\cdot 1 + 6\cdot 2\right) = \frac{33}{10} = 3{,}3 .$$

$$\overline{X^2} = \frac{1}{10}\left(1^2\cdot 1 + 2^2\cdot 4 + 3^2\cdot 1 + 4^2\cdot 1 + 5^2\cdot 1 + 6^2\cdot 2\right) = \frac{139}{10} = 13{,}9 ;$$

$$\sigma^{2*} = \overline{X^2} - \left(\overline{X}\right)^2 = 13{,}9 - 3{,}3^2 = 13{,}9 - 10{,}89 = 3{,}01 ;$$

$$s^2 = \frac{10}{9}\sigma^{2*} = \frac{10\cdot 3{,}01}{9} = \frac{301}{90} \approx 3{,}344 ; \qquad s \approx \sqrt{3{,}344} \approx 1{,}829 .$$

Пример 22 (продолжение). Добавим в таблицу столбец, содержащий значения середин интервалов $\overline{x_i}$ и вычислим статистики.

Если интервал является неограниченным, то для вычисления середины считаем, что длина этого интервала равна длине соседнего.

i	$[x_{i-1}, x_i)$	n_i	$\overline{x_i}$
1	95–100 %	8	97,5
2	100–105 %	28	102,5
3	105–110 %	36	107,5
4	110–115 %	24	112,5
5	больше 115 %	4	117,5

$$\overline{X} = \frac{1}{100}(97{,}5 \cdot 8 + 102{,}5 \cdot 28 + 107{,}5 \cdot 36 + 112{,}5 \cdot 24 + 117{,}5 \cdot 4) = \frac{10690}{100} = 106{,}9;$$

$$\overline{X^2} = \frac{1}{100}\left(97{,}5^2 \cdot 8 + 102{,}5^2 \cdot 28 + 107{,}5^2 \cdot 36 + 112{,}5^2 \cdot 24 + 117{,}5^2 \cdot 4\right) = 11452{,}25;$$

$$\sigma^{2*} = 11452{,}25 - 106{,}9^2 = 24{,}64\text{; откуда } s^2 = \frac{100}{99} 24{,}64 \approx 24{,}889.$$

4) ***Выборочная медиана***: $\tilde{Me}$ – середина вариационного ряда. Если N – четное, то $\tilde{Me}$ – среднее арифметическое двух средних элементов.

5) ***Выборочная мода***: $\tilde{Mo}$ – варианта, которой соответствует наибольшая частота.

6) ***Вариационный размах***: $R = X_{(N)} - X_{(1)}$.

Пример 21 (продолжение). Вариационный ряд:

1; 2; 2; 2; **<u>2; 3;</u>** 4; 5; 6; 6.

$$\tilde{Me} = \frac{X_{(5)} + X_{(6)}}{2} = \frac{2+3}{2} = 2{,}5;\ \tilde{Mo} = 2;\ R = X_{(10)} - X_{(1)} = 6 - 1 = 5.$$

6. ТОЧЕЧНОЕ ОЦЕНИВАНИЕ ПАРАМЕТРОВ РАСПРЕДЕЛЕНИЯ

6.1. Статистические оценки и их свойства

Пусть $X_1, X_2, ..., X_N$ – выборка объема N из известного распределения с функцией распределения $F(x, \theta)$ и неизвестным параметром θ.

Оценкой θ^* неизвестного параметра θ называется функция от выборки (статистика) $\theta^* = \theta^*(X_1, X_2, ..., X_N)$, посредством которой судят о значении параметра θ.

Задачей математической статистики является построение статистики θ^* так, чтобы она оценивала θ как можно лучше. Качество оценки определяется двумя свойствами – *несмещенностью* и *состоятельностью*.

Будучи функцией выборки, θ^* является случайной величиной, которая, возможно, имеет математическое ожидание и дисперсию.

Оценка θ^* называется ***несмещенной***, если $M\theta^* = \theta$.

Если равенство не выполнено, то оценка имеет смещение, и приближенные значения параметра, получаемые с ее помощью, будут либо заниженными (если $M\theta^* < \theta$), либо завышенными (если $M\theta^* > \theta$).

Оценка θ^* называется ***состоятельной***, если $\lim\limits_{N \to \infty} P\left(\left|\theta^* - \theta\right| > \varepsilon\right) = 0$ для любого $\varepsilon > 0$. Фактически состоятельность оценки означает, что $\theta^* \to \theta$ при $N \to \infty$; то есть с ростом объема выборки качество оценки растет, а значение асимптотически приближается к теоретическому значению параметра.

Теорема 13. Если θ^* – несмещенная оценка неизвестного параметра θ (т. е. $M\theta^* = \theta$) и $D\theta^* \to 0$ при $N \to \infty$, то θ^* – состоятельная оценка.

Более важным среди рассматриваемых двух свойств является состоятельность. Если состоятельная оценка имеет смещение, то можно внести поправку, которая это смещение устранит. Но если оценка несостоятельна, то есть ее качество не улучшается с ростом выборки, то использовать ее нельзя [3].

6.2. Оценки математического ожидания и дисперсии

Пусть $X_1, X_2, ..., X_N$ – выборка объема N из некоторого распределения, имеющего математическое ожидание и дисперсию. Пусть $MX_i = a$, $DX_i = \sigma^2$ для всех $i = 1, 2, ..., N$. Тогда

1) Выборочное среднее $\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$ является состоятельной и несмещенной оценкой математического ожидания.

$$M\overline{X} = M\left(\frac{1}{N}\sum_{i=1}^{N} X_i\right) = \frac{1}{N}M\left(\sum_{i=1}^{N} X_i\right) = \frac{1}{N}\sum_{i=1}^{N} MX_i = \frac{1}{N}Na = a.$$

В соответствии с определением оценка является несмещенной.

$$D\overline{X} = D\left(\frac{1}{N}\sum_{i=1}^{N} X_i\right) = \frac{1}{N^2}D\left(\sum_{i=1}^{N} X_i\right) = \left|\text{так как } X_i \text{ независимы}\right| = \frac{1}{N^2}\sum_{i=1}^{N} DX_i =$$

$$= \frac{1}{N^2}\left(\sigma^2 + \sigma^2 + ... + \sigma^2\right) = \frac{1}{N^2}N\sigma^2 = \frac{\sigma^2}{N} \to 0 \text{ при } N \to \infty .$$

В соответствии с теоремой 13 оценка состоятельна.

2) Выборочная дисперсия $\sigma^{2*} = \frac{1}{N}\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2$ является состоятельной оценкой, но имеет смещение.

3) Исправленная выборочная дисперсия $s^2 = \frac{N}{N-1}\sigma^{2*}$ – состоятельная и несмещенная оценка дисперсии.

При большом объеме выборки смещение выборочной дисперсии является незначительным, и поправку можно не вносить. Если объем выборки мал, следует пользоваться несмещенной оценкой дисперсии.

6.3. Оценка параметров распределения методом моментов

<u>Случай, когда неизвестен один параметр</u>

Пусть $X_1, X_2, ..., X_N$ – выборка объема N из известного распределения с функцией распределения $F(x,\theta)$, где θ – неизвестный параметр.

Зная функцию распределения, можно вычислить среднее элементов выборки, которое будет зависеть от параметра θ, $MX_i = M(\theta)$. Так как $\overline{X}$ – состоятельная и несмещенная оценка среднего, полагая $\overline{X} = M(\theta^*)$, можно найти θ^* – оценку θ.

Получим оценки параметров некоторых распределений.

Распределение Бернулли. Пусть дана выборка из распределения Бернулли $X_1, X_2, ..., X_N$. Требуется оценить параметр p. В п. 2.6. было вычислено:

$$MX_i = p, \text{ откуда } \overline{X} = p^*, \text{ а } \boxed{p^* = \overline{X}.}$$

Далее в случае каждого закона распределения используем формулы для вычисления математического ожидания, полученные в п. 2.6.

Биномиальное распределение при известном n**.** Требуется оценить параметр p.

$$MX_i = np, \text{ откуда } \overline{X} = np^*, \text{ и } \boxed{p^* = \frac{\overline{X}}{n}.}$$

Распределение Пуассона. Требуется оценить параметр λ.

$$MX_i = \lambda, \text{ откуда } \overline{X} = \lambda^*, \text{ и } \boxed{\lambda^* = \overline{X}.}$$

Геометрическое распределение. Требуется оценить параметр p.

$$MX_i = \frac{1}{p}, \text{ откуда } \overline{X} = \frac{1}{p^*}, \text{ и } \boxed{p^* = \frac{1}{\overline{X}}.}$$

Показательное распределение. Требуется оценить параметр λ.

$$MX_i = \frac{1}{\lambda}, \text{ откуда } \overline{X} = \frac{1}{\lambda^*}, \text{ и } \boxed{\lambda^* = \frac{1}{\overline{X}}.}$$

Равномерное распределение на отрезке $\left[0;d\right]$. Оценить параметр d .

$$MX_i = \frac{d+0}{2} = \frac{d}{2}, \text{ откуда } \overline{X} = \frac{d^*}{2}, \text{ и } \boxed{d^* = 2\overline{X}.}$$

<u>Случай, когда неизвестны два параметра</u>

Пусть $X_1, X_2, ..., X_N$ – выборка объема N из известного распределения с функцией распределения $F\left(x, \theta_1, \theta_2\right)$, где θ_1 и θ_2– неизвестные параметры.

Зная закон распределения, можно вычислить математическое ожидание и дисперсию элементов выборки, которые будут зависеть от параметров θ_1 и θ_2, $MX_i = M\left(\theta_1, \theta_2\right)$, $DX_i = D\left(\theta_1, \theta_2\right)$. Так как $\overline{X}$ и s^2 – состоятельные и несмещенные оценки математического ожидания и дисперсии, то θ_1^* и θ_2^* можно найти, решив систему уравнений:

$$\begin{cases} \overline{X} = M\left(\theta_1^*, \theta_2^*\right); \\ s^2 = D\left(\theta_1^*, \theta_2^*\right). \end{cases}$$

Получим оценки параметров некоторых распределений, как и ранее, используя формулы для вычисления математического ожидания и дисперсии каждого из распределений из п. 2.6.

Нормальное распределение. Требуется оценить параметры a и σ^2 .

$$MX_i = a,\ DX_i = \sigma^2, \text{ откуда } \overline{X} = a^*,\ s^2 = \left(\sigma^2\right)^*.$$

Таким образом, $\boxed{a^* = \overline{X},\ \left(\sigma^2\right)^* = s^2.}$

Оценка параметра σ^2 обозначается $\left(\sigma^2\right)^*$, и ее не следует путать с выборочной дисперсией σ^{2*} (смещенной оценкой).

Равномерное распределение на отрезке $\left[a;b\right]$. Требуется оценить параметры a и b .

$$MX_i = \frac{b+a}{2},\ DX_i = \frac{\left(b-a\right)^2}{12}, \text{ откуда } \overline{X} = \frac{b^*+a^*}{2},\ s^2 = \frac{\left(b^*-a^*\right)^2}{12}.$$

$$\begin{cases} b^* + a^* = 2\overline{X} \\ \left(b^* - a^*\right)^2 = 12s^2 \end{cases} \Rightarrow \begin{cases} b^* + a^* = 2\overline{X} \\ b^* - a^* = 2\sqrt{3s^2} \end{cases}.$$

Решив систему уравнений, получим: $\boxed{a^* = \overline{X} - \sqrt{3s^2},\ b^* = \overline{X} + \sqrt{3s^2}.}$

Биномиальное распределение. Требуется оценить параметры n и p.

$MX_i = np$, $DX_i = np(1-p)$, следовательно $\begin{cases} \overline{X} = n^* p^* \\ s^2 = n^* p^* \left(1 - p^*\right) \end{cases}$, откуда

$$\boxed{p^* = 1 - \frac{s^2}{\overline{X}},\ n^* = \frac{\overline{X}}{p^*}.}$$

В качестве оценки числа испытаний берется ближайшее целое число. Значение оценки p^* должно лежать в интервале $(0;1)$. Если это не так, то либо выборка не является репрезентативной, либо закон распределения ошибочно идентифицирован как биномиальный.

7. ОСНОВНЫЕ РАСПРЕДЕЛЕНИЯ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

Рассмотрим несколько законов распределения, которые в дальнейшем будут использованы для построения статистических оценок и критериев проверки статистических гипотез.

I. Нормальное распределение – N_{a,σ^2}

Нормальный закон является важнейшим из распределений теории вероятностей и математической статистики. Случайные величины, представленные в этом разделе, являются функциями нормально распределенных величин [3].

Из п. 2.6. известны плотность нормального распределения и его числовые характеристики:

$$\varphi_{a,\sigma^2}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-a)^2}{2\sigma^2}}, \text{ где } x \in (-\infty; +\infty),\ MX = a;\ DX = \sigma^2.$$

Стандартное нормальное распределение $N_{0,1}$:

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}; \qquad MX = 0;\ DX = 1.$$

II. Распределение Пирсона («хи-квадрат») – H_N

Пусть независимые случайные величины $X_1, X_2, ..., X_N$ распределены по стандартному нормальному закону $N_{0,1}$. Составим случайную величину: $\chi_N^2 = X_1{}^2 + X_2{}^2 + ... + X_N{}^2$. Говорят, что χ_N^2 имеет распределение ***«хи-квадрат» с*** N ***степенями свободы,*** или ***распределение Пирсона***.

Функция $h_N(x)$ – плотность распределения случайной величины χ_N^2 изображена на рис. 19. В табл. 5 в приложении к сборнику задач [4] даны значения обратного распределения Пирсона. В табл. 5 при известном числе

степеней свободы N и заданной вероятности α можно найти число $\chi^2_{\alpha,N}$ такое, что $P\left(\chi^2_N > \chi^2_{\alpha,N}\right) = \int\limits_{\chi^2_{\alpha,N}}^{+\infty} h_N(x)\,dx = \alpha$ (рис. 20).

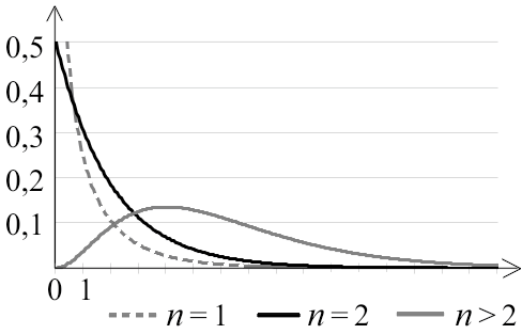


Рис. 19. График функции $h_N(x)$

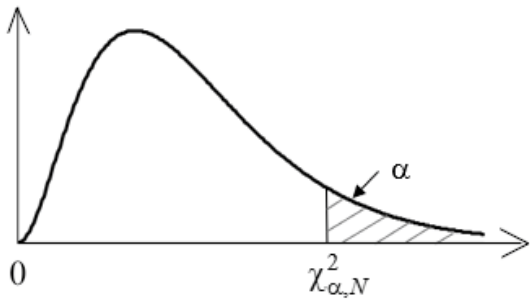


Рис. 20. Пояснение к табл. 5

III. Распределение Стьюдента («*t*-распределение») – T_N

Пусть независимые случайные величины $X_0, X_1, X_2, ..., X_N$ распределены по стандартному нормальному закону $N_{0,1}$. Составим случайную величину:

$$t_N = \frac{X_0}{\sqrt{\frac{1}{N}\left(X_1{}^2 + X_2{}^2 + ... + X_N{}^2\right)}} = \frac{X_0}{\sqrt{\frac{1}{N}\chi^2_N}}\ .$$

Говорят, что случайная величина t_N ***имеет t-распределение с*** N ***степенями свободы,*** или ***распределение Стьюдента***.

Функция $t_N(x)$ – плотность распределения случайной величины t_N в сравнении с плотностью стандартного нормального распределения изображена на рис. 21. В табл. 4 [4] даны значения обратного распределения Стьюдента, где при известном числе степеней свободы N и заданной вероятности γ можно найти число $t_{\gamma,N}$ такое, что $P\left(\left|t_N\right| < t_{\gamma,N}\right) = \int\limits_{-t_{\gamma,N}}^{t_{\gamma,N}} t_N(x)\,dx = \gamma$ (рис. 22).

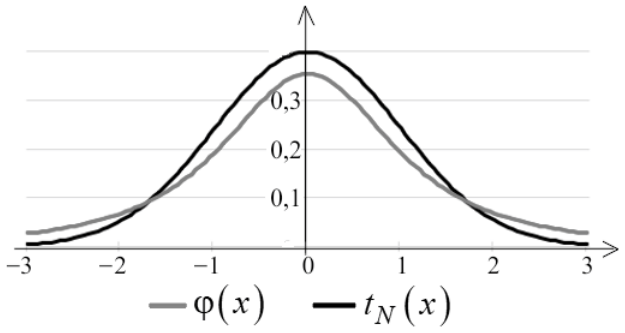


Рис. 21. График функции $t_N(x)$

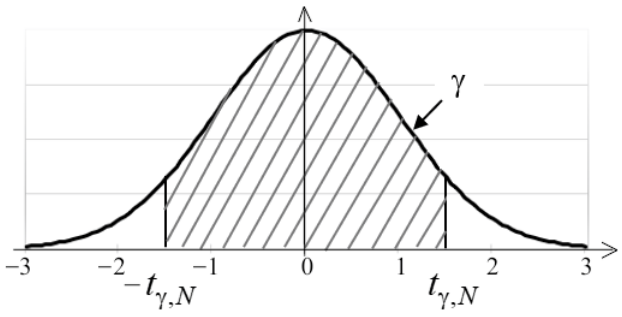


Рис. 22. Пояснение к табл. 4

8. ИНТЕРВАЛЬНОЕ ОЦЕНИВАНИЕ ПАРАМЕТРОВ РАСПРЕДЕЛЕНИЯ

8.1. Понятие доверительного интервала

Пусть $X_1, X_2, ..., X_N$ – выборка объема N из известного распределения с функцией распределения $F(x, \theta)$, где θ – неизвестный параметр. Пусть также задано число ε, $0 < \varepsilon < 1$. Интервал $(\theta^-; \theta^+)$ называется ***точным доверительным интервалом уровня*** $1 - \varepsilon$ для параметра θ, если для любого θ верно

$$P(\theta^- < \theta < \theta^+) \geq 1 - \varepsilon.$$

Величина $1 - \varepsilon$ называется ***уровнем доверия*** и обычно берется близкой к 1, например, 0,9; 0,95; 0,99 и т. п.

По возможности следует стремиться достичь равенства в условии, описывающем доверительный интервал, то есть $P(\theta^- < \theta < \theta^+) = 1 - \varepsilon$, так как в этом случае уровень доверия будет максимально возможным. Можно сказать, что доверительный интервал – это интервал, в который попадает теоретическое значение параметра θ с заданной вероятностью $1 - \varepsilon$.

Очевидно, что качественная, полезная оценка предполагает малую длину интервала и высокий уровень доверия. Однако при увеличении уровня доверия интервал становится шире. В частности, интервалом уровня 1 для любого параметра является $(-\infty; +\infty)$. Соответственно, при сужении интервала уровень доверия уменьшается.

Далее будут представлены методы построения точных доверительных интервалов для параметров нормального распределения.

8.2. Свойства выборок из нормального распределения

Рассмотрим следующие числовые характеристики выборки:

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i\,; \qquad s^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2;$$

Пусть известно, что $MX_i = a$. Статистикой, являющейся оценкой дисперсии при известном математическом ожидании, будет

$$\tilde{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}\left(X_i - a\right)^2 = \frac{1}{N}\sum_{i=1}^{N}\left(X_i^2 - 2aX_i + a^2\right) = \overline{X^2} - 2a\overline{X} + a^2.$$

Лемма Фишера. Пусть $X_1, X_2, ..., X_N$ – выборка из $N\left(a, \sigma^2\right)$. Тогда

1) случайная величина $\sqrt{N}\,\dfrac{\overline{X} - a}{\sigma}$ имеет распределение $N_{0,1}$;

2) случайная величина $\dfrac{N\tilde{\sigma}^2}{\sigma^2}$ имеет распределение H_N;

3) случайная величина $\dfrac{(N-1)s^2}{\sigma^2}$ имеет распределение H_{N-1};

4) случайная величина $\sqrt{N}\,\dfrac{\overline{X} - a}{\sqrt{s^2}}$ имеет распределение T_{N-1}.

8.3. Доверительные интервалы для параметров нормального распределения

Нормальное распределение N_{a,σ^2} имеет два параметра. Возможна ситуация, когда один из параметров известен, а другой необходимо оценить, а также когда неизвестны оба параметра. Получим формулы для каждого из случаев.

1) Для a при известном σ^2.

Из п. **1)** леммы Фишера известно, что случайная величина $\sqrt{N}\,\dfrac{\overline{X} - a}{\sigma}$ имеет распределение $N_{0,1}$. Это означает, что можно найти $\tau > 0$ такое, что

$$P\left(-\tau<\sqrt{N}\,\frac{\overline{X}-a}{\sigma}<\tau\right)=1-\varepsilon\,.$$

То есть, необходимо указать интервал $(-\tau;\tau)$, в который случайная величина, распределенная по стандартному нормальному закону, попадает с вероятностью $1-\varepsilon$ (рис. 23). Значение τ находится из соотношения $\widetilde{\Phi}(\tau)=\frac{1-\varepsilon}{2}$, где $\widetilde{\Phi}(\tau)$ – интеграл Лапласа (табл. 3 [4]).

После того, как найдено значение τ, необходимо преобразовать неравенство в скобках и получить интервал, в котором лежит a.

$$1-\varepsilon=P\left(-\tau<\sqrt{N}\,\frac{\overline{X}-a}{\sigma}<\tau\right)=P\left(\overline{X}-\frac{\tau\sigma}{\sqrt{N}}<a<\overline{X}+\frac{\tau\sigma}{\sqrt{N}}\right).$$

Таким образом, значение a с вероятностью $1-\varepsilon$ лежит в интервале

$$\boxed{\left(a^{-};a^{+}\right)=\left(\overline{X}-\frac{\tau\sigma}{\sqrt{N}};\overline{X}+\frac{\tau\sigma}{\sqrt{N}}\right).}$$

2) Для σ^2 при известном a.

Из п. **2)** леммы Фишера случайная величина $\frac{N\tilde{\sigma}^2}{\sigma^2}$ имеет распределение H_N. Это означает, что можно подобрать числа g^- и g^+ такие, что $\int\limits_{g^+}^{+\infty}h_N(x)dx=\frac{\varepsilon}{2}$ и $\int\limits_{g^-}^{+\infty}h_N(x)dx=1-\frac{\varepsilon}{2}$ (рис. 24).

В табл. 5 [4] для нахождения g^+ берется $\alpha=\frac{\varepsilon}{2}$, для g^- – $\alpha=1-\frac{\varepsilon}{2}$. Число степеней свободы в данном случае равно N. Найденные значения удовлетворяют условию $P\left(g^-<\frac{N\tilde{\sigma}^2}{\sigma^2}<g^+\right)=1-\varepsilon$. Преобразуем неравенство в скобках, чтобы найти интервал, в котором лежит σ^2:

$$1-\varepsilon=P\left(g^-<\frac{N\tilde{\sigma}^2}{\sigma^2}<g^+\right)=P\left(\frac{g^-}{N\tilde{\sigma}^2}<\frac{1}{\sigma^2}<\frac{g^+}{N\tilde{\sigma}^2}\right)=P\left(\frac{N\tilde{\sigma}^2}{g^-}>\sigma^2>\frac{N\tilde{\sigma}^2}{g^+}\right).$$

В итоге значение σ^2 с вероятностью $1-\varepsilon$ находится в интервале

$$\boxed{\left(\sigma^{2-};\sigma^{2+}\right)=\left(\frac{N\tilde{\sigma}^2}{g^+};\frac{N\tilde{\sigma}^2}{g^-}\right).}$$

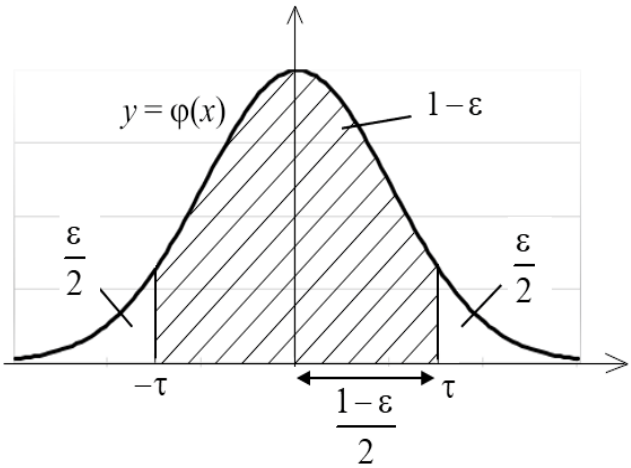


Рис. 23. Значение величины τ

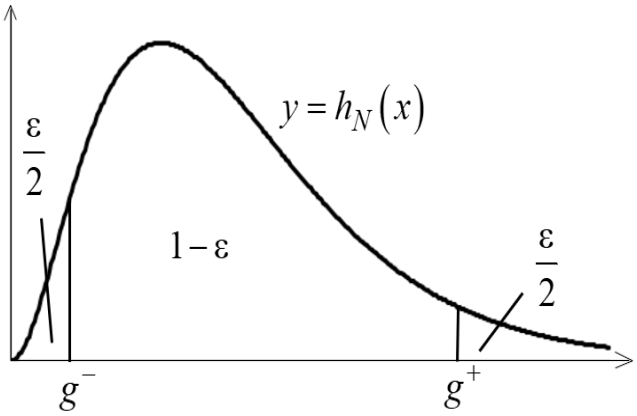


Рис. 24. Значение величин g^+ и g^-

3) Для σ^2 при неизвестном a.

Из п. **3)** леммы Фишера случайная величина $\frac{(N-1)s^2}{\sigma^2}$ имеет распределение H_{N-1}. Так же как и ранее, по табл. 5 [4] найдем числа g^+, соответствующее вероятности $\alpha=\frac{\varepsilon}{2}$, и g^- для $\alpha=1-\frac{\varepsilon}{2}$ при числе степеней свободы здесь равном $N-1$.

Рассуждая аналогично предыдущему случаю, получим доверительный интервал для σ^2 уровня $1-\varepsilon$:

$$\boxed{\left(\sigma^{2-};\sigma^{2+}\right)=\left(\frac{(N-1)s^2}{g^+};\frac{(N-1)s^2}{g^-}\right).}$$

4) Для a при неизвестном σ^2.

Из п. **4)** леммы Фишера случайная величина $\sqrt{N}\dfrac{\overline{X}-a}{\sqrt{s^2}}$ имеет распределение T_{N-1}. Можно подобрать число $t>0$ такое, что

$$P\left(-t<\sqrt{N}\frac{\overline{X}-a}{\sqrt{s^2}}<t\right)=1-\varepsilon.$$

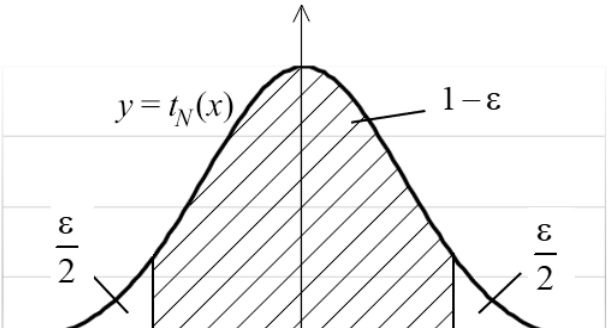


Рис. 25. Значение величины t

Случайная величина, имеющая распределение Стьюдента, попадает в интервал $(-t;t)$, с вероятностью $1-\varepsilon$ (рис. 25). Значение t ищется по табл. 4 [4], считая $\gamma=1-\varepsilon$, а число степеней свободы равным $N-1$.

Преобразовав неравенство в скобках, несложно найти интервал, в котором лежит a с вероятностью $1-\varepsilon$ лежит в интервале

$$\boxed{\left(a^-;a^+\right)=\left(\overline{X}-t\sqrt{\frac{s^2}{N}};\overline{X}+t\sqrt{\frac{s^2}{N}}\right)}.$$

Пример 23. Дана выборка из нормального распределения объема 25, для которой вычислены $\overline{X}=2,25$ и $\overline{X^2}=6,2025$. Найти точные доверительный интервал уровня 0,95 для параметров распределения.

Неизвестны оба параметра, a и σ^2 (случаи 3 и 4).

Из условия $N=25$; $1-\varepsilon=0,95$, откуда $\varepsilon=0,05$.

Кроме того, $\overline{X}=2,25$ и $\overline{X^2}=6,2025$, откуда

$$\sigma^{2*}=6,2025-2,25^2=1,14 \;\Rightarrow\; s^2=\frac{N}{N-1}\sigma^{2*}=\frac{25}{24}\cdot 1,14\overline{X}=2,25 \quad .$$

Для применения формул потребуются:

$$(N-1)s^2=25\cdot 1,14=28,5 \quad \text{и} \quad \frac{s^2}{N}=\frac{1,14}{24}=0,0475.$$

Найдем g^+ для $\alpha = \frac{\varepsilon}{2} = \frac{0,05}{2} = 0,025$ и g^- для $\alpha = 1 - \frac{\varepsilon}{2} = 0,975$. Число степеней свободы равно 24.

В табл. 5 за отсутствием нужных значений возьмем $\alpha = 0,02$, откуда $g^+ = 40,27$ и $\alpha = 0,98$, откуда $g^- = 11,99$. В результате интервал станет чуть шире и его уровень повысится до 0,96, что в принципе соответствует определению точного доверительного интервала.

Поставим найденные значения в формулу для случая 3:

$$\left(\sigma^{2-};\sigma^{2+}\right) = \left(\frac{28,5}{40,27};\frac{28,5}{11,99}\right) = \left(0,708;\ 2,377\right).$$

По табл. 4 для $\gamma = 1 - \varepsilon = 0,95$ и 24 степеней свободы находим $t = 2,064$. По формуле для случая 4:

$$\left(a^-;a^+\right) = \left(2,25 - 2,064\sqrt{0,0475};\ 2,25 - 2,064\sqrt{0,0475}\right) = \left(1,8;\ 2,7\right).$$

9. КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

9.1. Оценка коэффициента корреляции

Пусть дана выборка объема N: $(X_1;Y_1)$, $(X_2;Y_2)$, …, $(X_N;Y_N)$ из двумерного распределения. Оценкой теоретического коэффициента корреляции $r_{XY}=\dfrac{M\left[(X-MX)(Y-MY)\right]}{\sqrt{DX\cdot DY}}=\dfrac{M(XY)-MX\cdot MY}{\sqrt{DX\cdot DY}}$ является ***выборочный коэффициент корреляции***:

$$r^*_{XY}=\frac{\frac{1}{N}\sum\limits_{i=1}^{N}\left(X_i-\overline{X}\right)\left(Y_i-\overline{Y}\right)}{\sqrt{\sigma_X^{2*}\cdot\sigma_Y^{2*}}} \text{ или } r^*_{XY}=\frac{\overline{XY}-\overline{X}\cdot\overline{Y}}{\sqrt{\sigma_X^{2*}\cdot\sigma_Y^{2*}}}, \text{ где } \overline{XY}=\frac{1}{N}\sum_{i=1}^{N}X_iY_i\,.$$

Пример 24. При приеме на работу семи претендентам на должность предложили пройти два теста. Результаты тестов (в баллах) приведены в таблице. Оценить корреляцию между результатами тестов.

№	1	2	3	4	5	6	7
Тест 1, X	31	82	25	26	53	30	29
Тест 2, Y	21	55	8	27	32	42	26

Вычислим статистики отдельно для X и Y:

$$\overline{X}=\frac{1}{7}(31+82+25+26+53+30+29)=\frac{276}{7}\approx 39{,}429\,;$$

$$\overline{X^2}=\frac{1}{7}\left(31^2+82^2+25^2+26^2+53^2+30^2+29^2\right)=\frac{13536}{7}\approx 1933{,}714\,;$$

$$\sigma_X^{2*}=1933{,}714-39{,}429^2\approx 379{,}102\,.$$

Аналогично $\overline{Y}\approx 30{,}143$; $\sigma_Y^{2*}\approx 194{,}694$.

Вычислим $\overline{XY}$ и найдем выборочный коэффициент корреляции

$$\overline{XY} = \frac{1}{7}\left(31 \cdot 21 + 82 \cdot 55 + 25 \cdot 8 + 26 \cdot 27 + 53 \cdot 32 + 30 \cdot 42 + 29 \cdot 26\right) \approx 1396{,}143.$$

$$r_{XY}^{*} = \frac{\overline{XY} - \overline{X}\,\overline{Y}}{\sqrt{\sigma_X^{2*} \cdot \sigma_Y^{2*}}} = \frac{1396{,}143 - 39{,}429 \cdot 30{,}143}{\sqrt{379{,}102 \cdot 194{,}694}} \approx 0{,}764.$$

Как известно из п. 4.4, коэффициент корреляции принимает значения от –1 до 1. Чем ближе $\left|r_{XY}\right|$ или $\left|r_{XY}^{*}\right|$ к 1, тем более сильно выражена связь между X и Y; чем ближе r_{XY} (и r_{XY}^{*}) к нулю, тем менее ясен характер связи [3].

Значения коэффициента корреляции близкие к 1 возможны только если между X и Y существует функциональная линейная связь, то есть существуют числа a, b такие, что $Y = aX + b$.

Если элементы выборки $\left(X_1; Y_1\right)$, $\left(X_2; Y_2\right)$, …, $\left(X_N; Y_N\right)$ изобразить в виде точек на плоскости, то при низких значениях коэффициента корреляции картина может соответствовать рис. 26, при высоких – рис. 27.

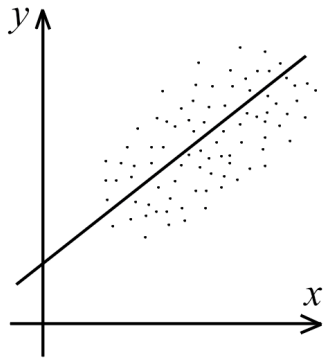


Рис. 26. Низкое значение r_{XY}^{*}

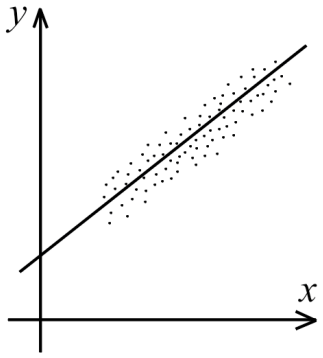


Рис. 27. Высокое значение r_{XY}^{*}

9.2. Уравнение линейной регрессии

Уравнение, описывающее функциональную зависимость между X и Y, называется ***уравнением регрессии*** Y по X. В частности, $Y = aX + b$ – ***уравнение линейной регрессии*** Y по X.

Уравнение линейной регрессии Y по X может быть записано в виде:

$$Y - \overline{Y} = l_Y\left(X - \overline{X}\right),$$

где $l_Y = \dfrac{\overline{XY} - \overline{X}\,\overline{Y}}{\sigma_X^{2*}} = r_{XY}^* \sqrt{\dfrac{\sigma_Y^{2*}}{\sigma_X^{2*}}}$ – ***коэффициент регрессии***.

Уравнение линейной регрессии может быть составлено и в том случае, когда коэффициент корреляции не близок к 1. Оно позволяет предсказать среднее ожидаемое значение Y при заданном X.

Можно также рассмотреть уравнение линейной регрессии X по Y, которое при заданном Y определяет среднее ожидаемое значение X. Для того чтобы получить его, поменяем местами переменные X и Y в формулах, приведенных выше:

$$X - \overline{X} = l_X\left(Y - \overline{Y}\right), \text{ где } l_X = \frac{\overline{XY} - \overline{X}\,\overline{Y}}{\sigma_Y^{2*}} = r_{XY}^* \sqrt{\frac{\sigma_X^{2*}}{\sigma_Y^{2*}}}.$$

При $r_{XY}^* = \pm 1$ эти уравнения задают одну и ту же прямую, однако если коэффициент корреляции отличен от 1, то прямые регрессии Y по X и X по Y и пересекаются в точке $\left(\overline{X};\overline{Y}\right)$ (рис. 28). Чем сильнее r_{XY}^* отличается от единицы, тем больше угол φ между прямыми.

Коэффициент корреляции позволяет выявить в первую очередь линейную между значениями случайных величин X и Y. Даже при близких к нулю значениях r_{XY}^* нельзя говорить об отсутствии связи между случайными величинами (рис. 29).

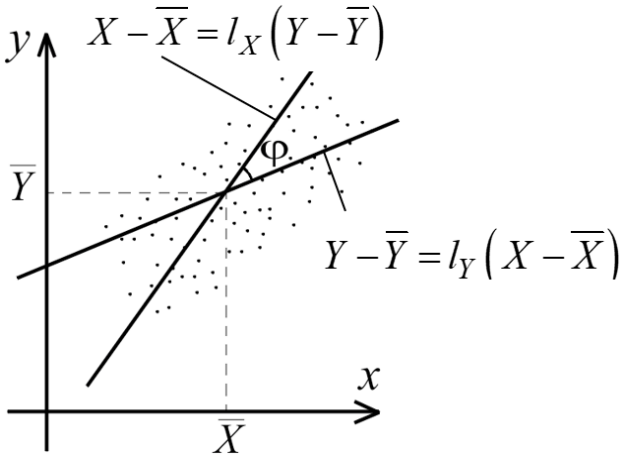


Рис. 28. Линейная регрессия

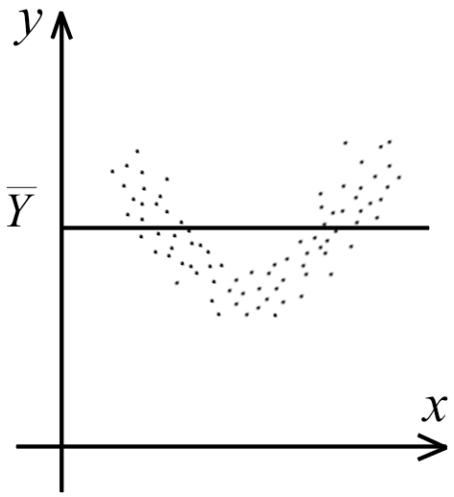


Рис. 29. Нелинейная регрессия

Пример 24 (продолжение). Построить прямую зависимости результатов второго теста от результатов первого. Сколько баллов, скорее всего, наберет соискатель во втором тесте, если в первом он набрал 90?

Найдем коэффициент регрессии:

$$l_Y = r_{XY}^* \sqrt{\frac{\sigma_Y^{2*}}{\sigma_X^{2*}}} = 0,764\sqrt{\frac{194,694}{379,102}} \approx 0,548.$$

Построим уравнение линейной регрессии Y по X: $Y - \overline{Y} = l_Y\left(X - \overline{X}\right)$.

$Y - 30,143 = 0,548\left(X - 39,429\right)$, откуда $Y = 0,548X - 0,548 \cdot 39,429 + 30,143$;

$$Y = 0,55X + 8,54;\ Y\left(90\right) = 0,55 \cdot 90 + 8,54 \approx 58.$$

В результате имеем: $Y = 0,55X + 8,54$, $Y\left(90\right) = 58$.

10. ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ

10.1. Статистические гипотезы и их проверка

Пусть дана выборка $X_1, X_2, ..., X_N$ объема N из генеральной совокупности, закон распределений которой либо неизвестен, либо известен и имеет функцию распределения $F(x, \theta)$ с неизвестным параметром θ.

Статистической гипотезой называется предположение о виде неизвестного распределения или о параметрах известного распределения.

Основная гипотеза, которая проверяется на основе материала выборки, называется ***нулевой*** и обозначается H_0. Одновременно с нулевой проверяется гипотеза H_1, которая противоречит нулевой и называется ***альтернативной,*** или ***конкурирующей***. Формулировка конкурирующей гипотезы определяется спецификой решаемой задачи.

Простой называют гипотезу, содержащую однозначное описание закона распределения. В противном случае гипотезу называют ***сложной***.

Пример 25. Пусть дана выборка из распределения Пуассона с неизвестным параметром λ . Нулевой гипотезой может быть $H_0 : \lambda = 2$. Альтернативной – $H_1 : \lambda = 3$ (простая) или $H_1 : \lambda \neq 2$ (сложная).

В ходе проверки необходимо, основываясь на анализе выборки, принять решение – считать справедливой нулевую гипотезу и отвергнуть альтернативную или, напротив, отвергнуть нулевую и принять конкурирующую [5]. Правило, по которому совершается выбор – принять или отвергнуть основную гипотезу, называется ***статистическим критерием***. Следует помнить, что решение, принятое в ходе проверки гипотезы, может оказаться ошибочным.

Ошибка первого рода состоит в том, что будет отвергнута правильная нулевая гипотеза. Вероятность ошибки первого рода называется ***уровнем значимости критерия*** и обозначается α.

Ошибка второго рода состоит в том, что будет отвергнута правильная альтернативная гипотеза. Вероятность ошибки второго рода обозначают β.

Мощностью критерия называют вероятность (1 – β), то есть вероятность принять альтернативную гипотезу в случае, если она верна.

В основе критерия всегда лежит работа со случайной величиной K, которая вычисляется по выборке и называется ***критической статистикой*** или ***критерием***, а ее значение для конкретной выборки – ***наблюдаемым значением критерия***. Наблюдаемое значение сравнивается с эталонной величиной $k_{\text{кр}}$ – ***критическим значением***, по результатам сравнения делается выбор в пользу основной или альтернативной гипотезы.

Критической областью называется совокупность значений критерия K, при которых H_0 отвергается. Существуют три вида критических областей.

Правосторонней называют критическую область, определяемую неравенством $K > k_{\text{кр}}$. В этом случае критическое значение выбирается так, чтобы было верно $P\left(K > k_{\text{кр}}\right) = \alpha$, где α – уровень значимости критерия.

Левосторонней называется критическая область, определяемая неравенством $K < k_{\text{кр}}$, где $k_{\text{кр}}$ такое, что для $P\left(K < k_{\text{кр}}\right) = \alpha$.

Двусторонней называется критическая область, определяемая неравенствами $K < k_1$, $K > k_2$, где $k_1 < k_2$. В частности, если критические точки симметричны относительно нуля, то двусторонняя критическая область определяется неравенством $|K| > k_{\text{кр}}$, где $k_{\text{кр}} > 0$. Для двусторонней симметричной области $P\left(K > k_{\text{кр}}\right) = \alpha/2$, $P\left(K < -k_{\text{кр}}\right) = \alpha/2$.

Задачи нахождения критической области и построения доверительного интервала связаны между собой. Существует также связь между видом альтернативной гипотезы и типом критической области [5].

Покажем это на следующих типовых постановках задач проверки статистических гипотез.

10.2. Проверка гипотезы о равенстве дисперсии нормально распределенной генеральной совокупности некоторому значению

Пусть дана выборка объема N из генеральной совокупности, распределенной по нормальному закону N_{a,σ^2}. Кроме того, есть основания

предполагать, что теоретическая дисперсия σ^2 равна некоторому значению σ_0^2. Выдвинем нулевую гипотезу $H_0{:}\sigma^2 = \sigma_0^2$ – «дисперсия генеральной совокупности равна предполагаемому значению σ_0^2».

Возможны три варианта альтернативной гипотезы:

1) $H_1{:}\sigma^2 > \sigma_0^2$; **2)** $H_1{:}\sigma^2 < \sigma_0^2$; **3)** $H_1{:}\sigma^2 \neq \sigma_0^2$.

По выборке можно вычислить исправленную выборочную дисперсию s^2. Необходимо проверить, значимо или незначимо различаются исправленная выборочная дисперсия s^2 и гипотетическая дисперсия σ_0^2.

Пусть задан уровень значимости α. Для проверки нулевой гипотезы рассмотрим критическую статистику – случайную величину $\chi^2 = \dfrac{(N-1)s^2}{\sigma_0^2}$, имеющую распределение «хи-квадрат» с $N-1$ степенями свободы (в силу п. 3) леммы Фишера, п. 9.2).

По заданным критерию и уровню значимости строится критическая область. Как уже говорилось ранее, вид критической области зависит от выбора конкурирующей гипотезы. Разберем возможные случаи.

1) $H_1{:}\sigma^2 > \sigma_0^2$. В этом случае критическая область строится как правосторонняя из условия $P\left(\chi^2 > \chi^2_{\text{кр}}\right) = \alpha$, а критическая точка находится по таблице распределения «хи-квадрат» (табл. 5 [4]) таким образом, что $\chi^2_{\text{кр}} = \chi^2_{\alpha,N-1}$. Далее по выборке вычисляется наблюдаемое значение критерия. Если $\chi^2 < \chi^2_{\text{кр}}$, то принимается нулевая гипотеза.

2) $H_1{:}\sigma^2 < \sigma_0^2$; В этом случае критическая область строится как левосторонняя. Критическая точка ищется как $\chi^2_{\text{кр}} = \chi^2_{1-\alpha,N-1}$. Тогда если $\chi^2 > \chi^2_{\text{кр}}$, то принимается нулевая гипотеза.

3) $H_1{:}\sigma^2 \neq \sigma_0^2$. В этом случае строится двусторонняя критическая область. Критические точки $\chi^2_{\text{пр}} = \chi^2_{\alpha/2,N-1}$ и $\chi^2_{\text{лев}} = \chi^2_{1-\alpha/2,N-1}$ находятся из условий $P\left(\chi^2 < \chi^2_{\text{лев}}\right) = \alpha/2$, $P\left(\chi^2 > \chi^2_{\text{прав}}\right) = 1-\alpha/2$ по табл. 5 [4].

Пример 26. При обработке выборки объема $N = 30$ получено значение $s^2 = 4$. Проверить гипотезу о том, что заданное значение $\sigma^2 = 5$ равно

дисперсии случайной величины, против двусторонней конкурирующей гипотезы на уровне значимости 0,1.

По условию задачи основная гипотеза $H_0{:}\sigma^2 = 5$; конкурирующая $H_1{:}\sigma^2 \neq 5$; $\alpha = 0{,}1$; откуда $\alpha/2 = 0{,}05$ и $1 - \alpha/2 = 0{,}95$. По таблице 5 [4] находим $\chi^2_{\text{лев}} = \chi^2_{0{,}05;29} = 17{,}71$ и $\chi^2_{\text{пр}} = \chi^2_{0{,}95;19} = 43{,}77$. Эти точки определяют границы двусторонней критической области $\left\{\chi^2 < 17{,}71 \text{ или } \chi^2 > 43{,}77\right\}$.

Вычислим наблюдаемое значение критерия:

$$\chi^2 = \frac{(N-1)s^2}{\sigma_0^2} = \frac{29 \cdot 4}{5} = 23{,}2.$$

Найденное значение не попадает в критическую область. Следовательно, принимается гипотеза H_0.

10.3. Проверка гипотезы о равенстве математического ожидания нормально распределенной генеральной совокупности некоторому значению

Пусть дана выборка объема N из генеральной совокупности, распределенной по нормальному закону N_{a,σ^2}. Кроме этого, есть основания предполагать, что генеральная средняя (математическое ожидание) равна некоторому значению a_0. Выдвинем нулевую гипотезу $H_0{:}\, a = a_0$ – «средняя данной генеральной совокупности равна значению a_0» при условии, что известен второй параметр – дисперсия σ^2.

Возможны три варианта альтернативной гипотезы:

1) $H_1{:}\, a \neq a_0$; **2)** $H_1{:}\, a > a_0$; **3)** $H_1{:}\, a < a_0$.

Требуется вычислить выборочное среднее $\overline{X}$ и при заданном уровне значимости α оценить, значимо или незначимо различаются $\overline{X}$ и a_0.

Для проверки нулевой гипотезы строится критерий – случайная величина $U = \sqrt{N}\,\dfrac{\overline{X} - a_0}{\sigma}$, которая имеет стандартное нормальное распределение $N_{0,1}$ в силу п. 1) леммы Фишера, п. 9.2.

Вид критической области зависит от выбора конкурирующей гипотезы. Разберем все возможные случаи.

1) Если $H_1: a \neq a_0$, то ищется двусторонняя критическая область с симметричными относительно нуля критическими точками, которые определяются из условия $\widetilde{\Phi}\left(u_{\text{кр}}\right) = \dfrac{1-\alpha}{2}$, где $\widetilde{\Phi}(x)$ − интеграл Лапласа.

2) Если $H_1: a > a_0$, то ищется правосторонняя критическая область с критической точкой, которая определяются из условия $\Phi\left(u_{\text{кр}}\right) = 1 - \alpha$, где $\Phi(x)$ − функция стандартного нормального распределения, или, что то же самое, $\widetilde{\Phi}\left(u_{\text{кр}}\right) = \dfrac{1}{2} - \alpha = \dfrac{1-2\alpha}{2}$, где $\widetilde{\Phi}(x)$ − интеграл Лапласа.

3) Если $H_1: a < a_0$, то ищется левосторонняя критическая область с критической точкой, которая определяются из условия $\widetilde{\Phi}\left(-u_{\text{кр}}\right) = \dfrac{1}{2} - \alpha$.

Пример 27. Проверить гипотезу о том, что заданное значение $a = 4$ равно математическому ожиданию нормально распределенной случайной величины с известной дисперсией, равной $\sigma^2 = 5$, против альтернативной двусторонней гипотезы на уровне значимости $\alpha = 0{,}05$, если в результате обработки выборки объема $N = 20$ получено $\overline{X} = 3$.

По условию задачи имеем: $H_0: a = 4$; $H_1: a \neq 4$; $\alpha = 0{,}05$; $(1-\alpha)/2 = 0{,}475$. Найдем критическое значение, для которого $\widetilde{\Phi}\left(u_{\text{кр}}\right) = 0{,}475$; по табл. 3 [4] $u_{\text{кр}} = 1{,}96$. Это точка, которая определяет двустороннюю критическую область $\{U < -1{,}96 \text{ или } U > 1{,}96\}$.

Вычислим наблюдаемое значение критерия $U = \sqrt{N}\dfrac{\overline{X} - a_0}{\sigma} = \sqrt{20}\dfrac{(3-4)}{\sqrt{5}} = -2$. Оно попадает в критическую область. Следовательно, гипотеза H_0 отклоняется.

Если второй параметр σ^2 неизвестен, то для проверки нулевой гипотезы выбирается другой критерий – случайная величина $T = \sqrt{N}\dfrac{\overline{X} - a}{\sqrt{s^2}}$,

которая имеет распределение Стьюдента с $N-1$ степенью свободы. Построим критические области для каждого из трех случаев.

1) Если $H_1: a \neq a_0$, то ищется двусторонняя критическая область с симметричными относительно нуля критическими точками, которые определяются из условия $t_{\text{кр}} = t_{1-\alpha, N-1}$ (раздел 8).

2) Если $H_1: a > a_0$, то ищется правосторонняя критическая область с критической точкой, которая определяется из условия $t_{\text{кр}} = t_{1-2\alpha, N-1}$.

3) Если $H_1: a < a_0$, то ищется левосторонняя критическая область с критической точкой, которая определяется из условия $t_{\text{кр}} = -t_{1-2\alpha, N-1}$.

Пример 28. Проверить гипотезу о том, что заданное значение $a = 10$ равно математическому ожиданию нормально распределенной случайной величины, против альтернативной правосторонней гипотезы на уровне значимости $\alpha = 0{,}05$, если в результате обработки выборки объема $N = 16$ получены статистики $\overline{X} = 12$ и $s^2 = 1$.

Имеем $H_0: a = 10$; $H_1: a > 10$; $\alpha = 0{,}05$; $1 - 2\alpha = 0{,}9$. Число степеней свободы равно 15. По табл. 4 [4] найдем $t_{\text{кр}} = t_{0{,}9;15} = 1{,}753$. Эта критическая точка определяет правостороннюю критическую область $T > 1{,}833$.

Наблюдаемое значение критерия $T = \sqrt{N}\dfrac{\overline{X} - a}{\sqrt{s^2}} = \sqrt{16}\dfrac{(12-10)}{1} = 8$ попало в критическую область; принимается альтернативная гипотеза H_1.

10.4. Проверка гипотезы о доле признака

Пусть дана выборка из распределения Бернулли объема N с неизвестной вероятностью успеха p, состоящая из m единиц и $N - m$ нулей, тогда $p^* = \overline{X} = \dfrac{m}{N}$ – оценка параметра p (п. 7.3).

Если $N > 50$, а выборочное значение p^* удовлетворяет условиям $Np^* > 5$ и $N\left(1 - p^*\right) > 5$, то для проверки гипотезы $H_0: p = p_0$ используют статистику

$$Z = \frac{m - Np}{\sqrt{Npq}} = \frac{p^* - p}{\sqrt{pq/N}}, \text{ где } q = 1 - p.$$

При условии, что гипотеза H_0 верна, критическая статистика Z имеет распределение, близкое к стандартному нормальному распределению $N_{0,1}$.

Критическая область критерия при уровне значимости α определяется условиями:

1) $|Z| < z_{\text{кр}}$ – при H_1: $p \neq p_0$, где $\widetilde{\Phi}(z_{\text{кр}}) = \frac{1-\alpha}{2}$, $\widetilde{\Phi}(x)$ – интеграл Лапласа;

2) $Z > z_{\text{кр}}$ – при H_1: $p > p_0$, где $\widetilde{\Phi}(z_{\text{кр}}) = \frac{1}{2} - \alpha$;

3) $Z < z_{\text{кр}}$ – при H_1: $p < p_0$, где $\widetilde{\Phi}(-z_{\text{кр}}) = \frac{1}{2} - \alpha$.

Пример 29. В партии из 600 изделий обнаружили 60 бракованных. Проверить гипотезу, что доля брака составляет 8 % при уровне значимости $\alpha = 0,05$ против альтернативной гипотезы, что брак больше 8 %.

По условию задачи имеем: $H_0: p = 0,08$; $H_1: p > 0,08$. Вычислим $p^* = \frac{60}{600} \approx 0,1$. Далее, $\alpha = 0,05$, $0,5 - \alpha = 0,45$. Найдем критическое значение, из условия $\widetilde{\Phi}(z_{\text{кр}}) = 0,45$. По табл. 3, $z_{\text{кр}} = 1,65$.

Вычислим наблюдаемое значение критерия:

$$Z = \frac{p^* - p}{\sqrt{pq/N}} \approx \frac{0,1 - 0,08}{\sqrt{0,08 \cdot 0,92/600}} \approx 0,18 < 1,65.$$

Наблюдаемое значение не попадает в критическую область, следовательно, считаем, что верна гипотеза H_0. Проверка показывает, что нет отличия доли брака от 8 % на уровне значимости 5 %.

10.5. Проверка гипотезы о равенстве средних двух нормально распределенных генеральных совокупностей

Пусть заданы две нормально распределенные генеральные совокупности. Рассмотрим случаи проверки гипотезы о равенстве математических ожиданий этих совокупностей, $H_0: a_1 = a_2$.

Критическая статистика для проверки гипотезы о равенстве средних нормально распределенных генеральных совокупностей ***при известных дисперсиях*** имеет вид

$$Z = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}},$$

где N_1, N_2 – объемы выборок; $\overline{X}_1$, $\overline{X}_2$ – выборочные средние исследуемых генеральных совокупностей; σ_1^2, σ_2^2 – известные дисперсии генеральных совокупностей. Случайная величина Z при больших объемах выборок имеет распределение близкое к стандартному нормальному $N_{0,1}$.

Критическая область критерия при уровне значимости α будет такой:

1) $|Z| > z_{\text{кр}}$ – при $H_1: a_1 \neq a_2$, где $\widetilde{\Phi}\left(z_{\text{кр}}\right) = \frac{1-\alpha}{2}$;

2) $Z > z_{\text{кр}}$ – при $H_1: a_1 > a_2$, где $\widetilde{\Phi}\left(z_{\text{кр}}\right) = \frac{1}{2} - \alpha$;

3) $Z < z_{\text{кр}}$ – при $H_1: a_1 < a_2$, где $\widetilde{\Phi}\left(-z_{\text{кр}}\right) = \frac{1}{2} - \alpha$;

Пример 30. При обработке выборок объемов $N_1 = 40$, и $N_2 = 20$ из генеральных совокупностей с известными дисперсиями $\sigma_1^2 = 4$, $\sigma_2^2 = 1$ получены выборочные средние $\overline{X}_1 = 4$, $\overline{X}_2 = 6$. На уровне значимости 5 % проверить гипотезу о равенстве математических ожиданий против конкурирующей гипотезы о том, что $a_1 < a_2$.

По условию задачи имеем: $H_0: a_1 = a_2$; $H_1: a_1 < a_2$. Далее, $\alpha = 0,05$, откуда $0,5 - \alpha = 0,45$. Найдем критическое значение, удовлетворяющее условию $\widetilde{\Phi}\left(-z_{\text{кр}}\right) = 0,45$ (аргумент интеграла Лапласа, при котором он принимает значение равное 0,45). По табл. 3 $z_{\text{кр}} = -1,65$.

Вычислим наблюдаемое значение критерия:

$$Z = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{4-6}{\sqrt{\frac{4}{40} + \frac{1}{20}}} \approx -5,17 < z_{\text{кр}} = -1,65.$$

Значение критерия попадает в критическую область. Следовательно, принимается альтернативная гипотеза $H_1: a_1 < a_2$.

Критическая статистика для проверки гипотезы о равенстве средних нормально распределенных генеральных совокупностей ***при неизвестных, но равных дисперсиях*** имеет вид:

$$T = \frac{\overline{X}_1 - \overline{X}_2}{s\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}}, \text{ где } s = \sqrt{\frac{(N_1 - 1)s_1^2 + (N_2 - 1)s_2^2}{N_1 + N_2 - 2}},$$

где N_1, N_2 – объемы выборок; $\overline{X}_1$, $\overline{X}_2$ – выборочные средние исследуемых генеральных совокупностей; s_1^2, s_2^2 – исправленные выборочные дисперсии исследуемых генеральных совокупностей. Статистика T имеет распределение Стьюдента $T_{N_1+N_2-2}$.

Критическая область при уровне значимости α определяется так:

1) $|T| < t_{\text{кр}}$ – при альтернативной гипотезе $H_1: a_1 \neq a_2$, где критическое значение ищется по табл. 4 из условия $t_{\text{кр}} = t_{1-\alpha, N_1+N_2-2}$.

2) $T > t_{\text{кр}}$ – при $H_1: a_1 > a_2$, где $t_{\text{кр}} = t_{1-2\alpha, N_1+N_2-2}$.

3) $T < t_{\text{кр}}$ – при $H_1: a_1 < a_2$, где $t_{\text{кр}} = -t_{1-2\alpha, N_1+N_2-2}$.

ЗАКЛЮЧЕНИЕ

Теория вероятностей и математическая статистика играют важную роль в работе специалиста по информационной безопасности, помогая изучать возможные угрозы и уязвимости систем и принимать обоснованные решения для обеспечения безопасности информации.

Данное учебное пособие дает обучающимся возможность освоить теоретический материал и методы решения практических задач курса дисциплины «Теория вероятностей и математическая статистика», без которых невозможно овладение инструментами и технологиями, необходимыми для понимания, проектирования и защиты информационных систем.

Знания, приобретаемые обучающимися в ходе изучения дисциплины «Теория вероятностей и математическая статистика», обязательны для дальнейшего формирования профессиональных компетенций при освоении таких специализированных дисциплин, как «Методы и средства криптографической защиты информации», «Системы искусственного интеллекта», «Теория информации», «Управление рисками информационных систем», «Теория и практика изменений» и др.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Бочаров П. П., Печинкин А. В. Теория вероятностей. Математическая статистика. – М. : Физматлит, 2005. – 296 с.

2. Гмурман В. Е. Теория вероятностей и математическая статистика : учеб. пособие для студентов вузов. – 12-е изд., перераб. – М. : ЮРАЙТ, 2009. – 478 с.

3. Кремер Н. Ш. Теория вероятностей и математическая статистика : учеб. – М. : Юнити, 2012. – 551 с.

4. Неклюдова В. Л., Вербная В. П. Теория вероятностей и математическая статистика : сб. задач. – Новосибирск : СГУГиТ, 2023. – 37 с.

5. Рапоцевич Е. А. Теория вероятностей и математическая статистика. – Новосибирск : СибАГС, 2012. − 94 с.

Учебное издание

Неклюдова Вера Леонидовна

Вербная Валентина Павловна

ТЕОРИЯ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Редактирование и компьютерная верстка
Ю. С. Мерзликиной

Изд. лиц. ЛР № 020461 от 04.03.1997.
Подписано в печать 28.05.2024. Формат 60 × 84 1/16.
Усл. печ. л. 5,34. Тираж 144 экз. Заказ 65.
Гигиеническое заключение
№ 54.НК.05.953.П.000147.12.02. от 10.12.2002.

Редакционно-издательский отдел СГУГиТ
630108, Новосибирск, ул. Плахотного, 10.

Отпечатано в картопечатной лаборатории СГУГиТ
630108, Новосибирск, ул. Плахотного, 8.